MODELING SEMANTIC AND ORTHOGRAPHIC SIMILARITY EFFECTS ON MEMORY FOR INDIVIDUAL WORDS

Mark Steyvers

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements for the degree
Doctor of Philosophy
in the Department of Psychology
Indiana University

September 2000

Abstract

Many memory models assume that the semantic and physical features of words can be represented by collections of features abstractly represented by vectors. Most of these memory models are process oriented; they explicate the processes that operate on memory representations without explicating the origin of the representations themselves; the different attributes of words are typically represented by random vectors that have no formal relationship to the words in our language. In Part I of this research, we develop Word Association Spaces (WAS) that capture aspects of the meaning of words. This vector representation is based on a statistical analysis of a large database of free association norms. In Part II, this representation along with a representation for the physical aspects of words such as orthography is combined with REM, a process model for memory. Three experiments are presented in which distractor similarity, the length of studied categories and the directionality of association between study and test words were varied. With only a few parameters, the REM model can account qualitatively for the results. Developing a representation incorporating features of actual words makes it possible to derive predictions for individual test words. We show that the moderate correlations between observed and predicted hit and false alarm rates for individual words are larger than can be explained by models that represent words by arbitrary features. In Part III, an experiment is presented that tests a prediction of REM: words with uncommon features should be better recognized than words with common features, even if the words are equated for word frequency.

Acknowledgments

First and foremost, I would like to thank Rich Shiffrin who has been a great advisor and mentor. His influence on this dissertation work has been substantial and his insistence on aiming for only the best scientific research will stay with me forever. Also, Rob Goldstone has been an integral part of my graduate career with our many collaborations and stimulating conversations. I would also like to acknowledge my collaborators Ken Malmberg and Joseph Stephens in the research presented in part III of the dissertation and Tom Busey who provided both ideas and encouragement of any project of shared interest. I would also like to thank Eric-Jan Wagenmakers, Rob Nosofsky, and Dan Maki for their support and many helpful discussions. Last but not least, my friends Peter Grünwald, Mischa Bonn, and Dave Huber have always been supportive and I can highly recommend going out with these guys.

Contact: Mark Steyvers at msteyver@psych.stanford.edu Stanford University. Building 420, Jordan Hall, Stanford, CA 94305-2130, Tel: (650) 725-5487, Fax: (650) 725-5699

Part I:
Creating Semantic Spaces for Words based on Free Association Norms

It has been proposed that various aspects of words can be represented by separate collections of features that code for temporal, spatial, frequency, modality, orthographic, acoustic, and associative aspects of the words (Anisfeld & Knapp, 1968; Bower, 1967; Herriot, 1974; Underwood, 1969; Wickens, 1972). In part I of this research, we will focus on the associative/semantic aspects of words.

A common assumption is that the meaning of a word can be represented by a vector which places a word in a multidimensional semantic space (Bower, 1967; Landauer & Dumais, 1997; Lund & Burgess, 1996; Morton, 1970; Norman, & Rumelhart, 1970; Osgood, Suci, & Tannenbaum, 1957; Underwood, 1969; Wickens, 1972). The main requirement of such spaces is that words that are similar in meaning should be represented by similar vectors. Representing words as vectors in a multidimensional space allows simple geometric operations such as the Euclidian distance or inner product to compute the semantic similarity between arbitrary pairs or groups of words. This makes it possible to make predictions about performance in psychological tasks where the semantic distance between pairs or groups of words is assumed to play a role.

The main goal of part I of this research is to introduce a new method for creating psychological spaces that is based on an analysis of a large free association database collected by Nelson, McEvoy, and Schreiber (1998) containing norms for first associates for over 5000 words. This method places over 5000 words in a psychological space that we will call Word Association Space (WAS).

We believe such a construct will be very useful in the modeling of episodic memory phenomena since it has been shown that associative structure of words plays a central role in recall (e.g. Bousfield, 1953; Cramer, 1968; Deese, 1959a,b, 1965; Jenkins, Mink, & Russell, 1958), cued recall (e.g. Nelson, Schreiber, & McEvoy, 1992) and priming (e.g. Canas, 1990; see also Neely, 1991). For example, Deese (1959a,b) found that the inter-item associative strength for the words on a study list can predict the number of words recalled, the number of intrusions, and the frequency with which certain words intrude.

In this paper, we will first introduce four methods to create semantic spaces. These are based on the semantic differential, multidimensional scaling on similarity ratings, LSA, and HAL. Then, we will introduce WAS, the approach of placing words in a high dimensional space by analyzing free association norms. The similarity and differences between WAS and free association norms are discussed. Two demonstrations are given that WAS is useful in predicting memory performance. First, we will show that the intrusion rates in free recall experiments observed in Deese (1959b) can be predicted on the basis of the similarity structure in the vector space. Second, we will show that WAS can predict to some degree the percentage of correctly recalled words in extra list cued recall tasks (Nelson & Schreiber, 1992; Nelson, Schreiber, & McEvoy, 1992; Nelson, McKinney, Gee, & Janczura, 1998; Nelson & Xu, 1995). We will contrast the predictions from WAS with predictions made by the LSA approach.

Methods to Construct Semantic Spaces

Semantic differential. This method was developed by Osgood, Suci, and Tannenbaum (1957). Words are rated on a set of bipolar rating scales. The bipolar rating scales are semantic scales defined by pairs of polar adjectives (e.g. "good-bad", "altruistic-egotistic", "hot-cold"). Each word that one wants to place in the semantic space is judged on these scales. If numbers are assigned from low to high for the left to right word of a bipolar pair, then the word "dictator" for example, might be judged high on the "good-bad", high on the "altruistic-egotistic" and neutral on the "hot-cold" scale. For each word, the ratings averaged over a large number of subjects define the coordinates of the word in the semantic space. Because semantically similar words are likely to receive similar ratings, they are likely to be located in similar regions of the semantic space. The advantage of the semantic differential method is the simplicity and intuitive appeal. The problem inherent to this approach is the arbitrariness in choosing the set of semantic scales as well as the number of semantic scales.

MDS on similarity ratings. In this method, participants rate the semantic similarity for pairs of words. Then, those similarity ratings can be subjected to multidimensional scaling analyses to derive vector representations in which similar vectors represent words similar in meaning (Caramazza, Hersch, & Torgerson, 1976; Rips, Shoben, & Smith, 1973; Schwartz & Humphreys, 1973). While this method is straightforward and has led to interesting applications (e.g. Caramazza et al; Romney et al., 1993.), it is clearly impractical for large number of words as the number of ratings that must be collected goes up quadratically with the number of stimuli.

Latent Semantic Analysis (LSA). A method to derive high-dimensional semantic spaces that does not rely on judgments by participants is Latent Semantic Analysis or LSA (Derweester, Dumais, Furnas, Landauer, & Harshman, 1990; Landauer &

Dumais, 1997; Landauer, Foltz, & Laham, 1998). The assumption Landauer and Dumais (1997) make is that similar words occur in similar contexts. A context can be defined by any connected set of text from a corpus such as an encyclopedia, or samples of texts from textbooks. For example, a textbook with a paragraph about "cats" might also mention "dogs", "fur", "pets" etc. This knowledge can be used to assume that "cats" and "dogs" are related in meaning. However, some words are clearly related in meaning such as "cats" and "felines" but they might never occur simultaneously in the same context. There might be indirect links between "cats" through its context words with "felines", i.e., the words share similar contexts. The technique of singular value decomposition (SVD) can be applied on the matrix of word-context co-occurrence statistics. This methods analyzes the direct and indirect relationships between words and contexts in the matrix based on simple matrix-algebraic operations. The result of the SVD analysis is a high dimensional space in which words that appear in similar contexts are placed in similar regions of the space. Landauer and Dumais (1997) applied the LSA approach on the 68,000 words of a large encyclopedia and placed these words in a high dimensional space with the number of dimensions chosen between 100 and 400. The LSA representation has been successfully applied to multiple choice vocabulary tests, domain knowledge tests and content evaluation (see Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998).

Hyperspace Analogue to Language (HAL). The HAL model develops high dimensional vector representations for words that like LSA is based on a co-occurrence analysis of large samples of written text (Burgess, Livesay, & Lund, 1998; Lund & Burgess, 1996; see Burgess & Lund, 2000 for an overview). For 70,000 words, the co-occurrence statistics were calculated in a 10 word window that was slid over the text from a corpus of over 320 million words (gathered from Usenet newsgroups). For each word, the co-occurrence statistics were calculated of the 70,000 words appearing before or after that word in the 10 word window. The resulting 140,000 values for each word were the feature values for the words in the HAL representation. Because the representation is based the context in which words appear, the HAL vector representation is also referred to as a contextual space: words that appear in similar contexts are represented by similar vectors. The HAL and LSA approach are similar in one major assumption: similar words occur in similar contexts. In both HAL and LSA, the placement of words in a high dimensional semantic space is based on an analysis of the co-occurrence statistics of words in their contexts. In LSA, a context is defined by a relatively large segment of text whereas in HAL, the context is defined by a window of 10 words[1].

One great advantage of LSA and HAL over approaches depending on human judgments is that almost any number of words can be placed in a semantic/contextual space. This is possible because the method relies uniquely on samples of written text (of which there is a virtually unlimited amount) as opposed to ratings provided by participants. Even though a working vocabulary of 5000 words in WAS is much smaller than the 70,000 word long vocabularies of LSA and HAL, we believe it is large enough for our purpose of modeling performance in memory tasks.

Word Association Spaces

Deese (1962,1965) asserted that free associations are not haphazard processes in our brain and that there is regularity underneath them. He laid the framework for studying the meaning of linguistic forms that can be derived by analyzing the correspondences between distributions of responses to free association stimuli: "The most important property of associations is their structure - their patterns of intercorrelations" (Deese, 1965, p.1). The SVD method has been successfully applied in LSA to uncover the patterns of intercorrelations of the co-occurrence statistics for words appearing in contexts. We will also use the SVD method but apply it on a different database: a large database of free association norms collected by Nelson, McEvoy, and Schreiber (1998) containing norms for first associates for over 5000 words.

In total, more than 6000 people participated in the collection of this database. An average of 149 (SD = 15) participants were presented with 100-120 English words. These words served as cues (e.g. "cat") for which participants had to write down the first word that came to mind (e.g. "dog"). These experiments were performed on many participants so that for each cue the relative associative strengths could be calculated for responses by the proportion of subjects that elicited the response to the cue (e.g. 60% responded with "dog", 15% with "pet", 10% with "tiger", etc).

The idea is to apply the SVD method to place words in a high dimensional space by analyzing the direct and indirect associative relationships between words. While the details of this procedure are discussed in the Appendix, the basic approach is illustrated in Figure 1. The free association norms were represented in matrix form. The rows represent the cues and the columns represent the responses. An entry in the matrix represents the relative frequency with which a response was generated for the particular cue (i.e., associative strength). Before SVD was applied to the matrix, it was preprocessed in two ways. First, the indirect associative strengths between words were calculated and added to the matrix[6]. Then, the matrix was symmetrized such that the associative strength between cue A and response B equaled the associative strength between cue B and response A. After these preprocessing steps, the matrix was subjected to SVD. The result of SVD is the placement of words in a high dimensional space, which we called Word Association Space (WAS).

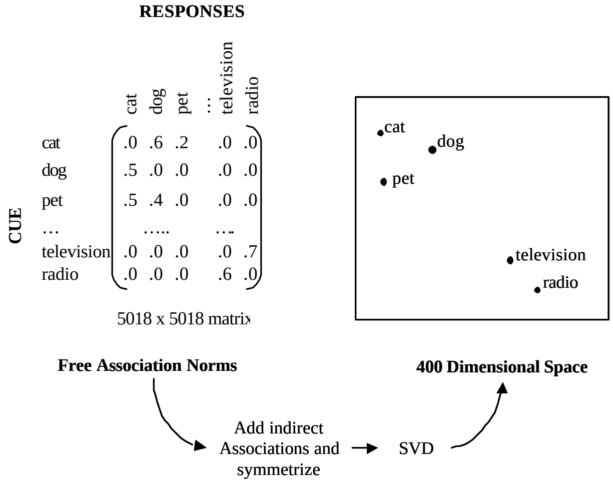


Figure 1. Illustration of the creation of Word Association Spaces (WAS). By singular value decomposition on a large database of free association norms, words are placed in a high dimensional semantic space. Words with similar associative relationships are placed in similar regions of the space.

In WAS, words that have similar associative structures are represented by similar vectors. Words that are not direct associates of each other can also be represented by similar vectors if their associates are related (or if the associates of the associates of the words are related).

The representation of words in WAS is dependent on the method with which the free association norms are analyzed. By using the SVD method, words are represented by vectors with continuous feature values that have a symmetric distribution around zero. A suitable measure for the similarity between two words is the inner product of the two word vectors. The idea is that two words that are similar in meaning or that have similar associative structures have high similarity as defined by the inner product of the two word vectors.

An important variable (which we will call k) is the number of dimensions of the space[2]. One can think of k as the number of feature values for the words. We vary k between 10 and 400. The number of dimensions will determine how much the information of the free association database is compressed. With too few dimensions, the similarity structure of the resulting vectors does not capture enough detail of the original associative structure in the database. With too many dimensions, the similarity structure of the vectors does not capture enough of the indirect relationships in the associations between words.

To get an understanding of what the similarity structure of WAS is like, we performed four analyses. In the first analysis, the similarity structure of low and high frequency is compared and it is shown that in WAS, high frequency words are more similar to other high frequency words than to low frequency words. In the second analysis, we compared the ordering of neighbors in WAS to the ordering of the strength of associates in the free association norms. In the third analysis, the issue of whether WAS captures semantic or associative relationships (or both) is addressed. It is argued that it is difficult to make a distinction between the two kinds of relationships. In the fourth analysis, we analyze the ability of WAS to capture the differences between and within semantic categories. We will now discuss these four analyses in turn.

Word Frequency and the Similarity Structure in WAS

Word frequency can be defined by the number of times words occur in large samples of written text (Kucera & Francis). The frequency of words in samples of written text correlates with the frequency with which words are produced in free association norms. High frequency words are produced more often as responses in free association norms[3]. We investigated the similarity structure of low and high frequency words in WAS by calculating the similarity between groups of words with different frequency ranges. In Figure 2, top panel, the average inner product is calculated between random words from different Kucera and Francis frequency ranges. The highest similarity was obtained between high frequency words. Lower similarities were obtained between high and low frequency words and the lowest similarity was obtained between low frequency words. The reason for the average similarity being higher between high frequency words is that high frequency word vectors in WAS have larger magnitudes than low frequency word vectors. This is shown in Figure 2, bottom panel. Vectors with larger magnitudes, on average lead to larger inner products.

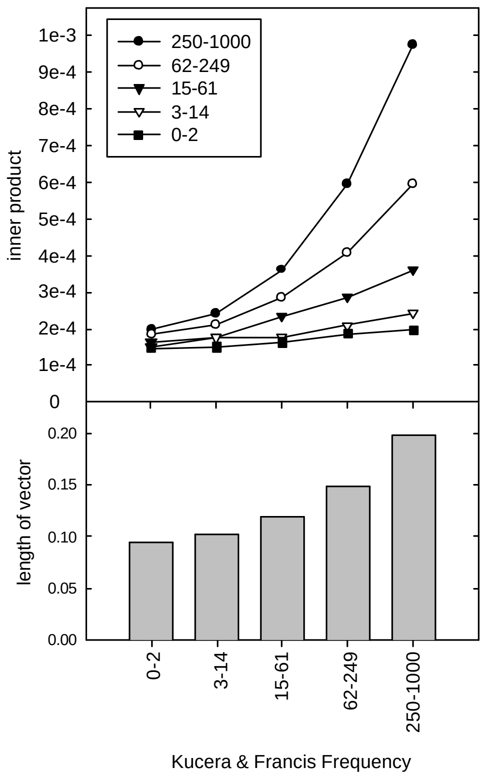


Figure 2. The effect of word frequency on the similarity structure of WAS and the length of the word vectors. In the top panel, the average similarity (measured by inner product) between random words from different Kucera and Francis word frequency ranges is plotted. The similarity is highest when high frequency words are compared with high frequency words.

The similarity decreases when the word frequencies of the words compared decreases. In the bottom panel, the figure shows that the vector lengths are bigger for high frequency words than low frequency words. Of course, it is the combination of the vector magnitudes and the correlation between the feature values that determine the similarity as computed by the inner product. Because high frequency words on average have larger magnitudes, they are placed more at the outskirts of the semantic space while low frequency words are placed more in the center of the space. Because an inner product measure for similarity is used, the average similarity between the high frequency words that lie at the outskirts of the space is higher than between words that lie more in the center of the space. Of course, using a different similarity measure should lead to different results. For example, using Euclidian distance as a measure for (inverse) similarity, should lead to lower similarities between high than low frequency words. This observation becomes important for part II of this research.

Predicting the Output Order of Free Association Norms

Because the word vectors in WAS are based explicitly on the free association norms, it is of interest to check whether the output order of responses (in terms of associative strength) can be predicted by WAS. We took the 10 strongest responses to each of the cues in the free association norms and ranked them according to associative strengths. For example, the response ‘crib’ is the 8th

Table 1
Median rank of the output-order in WAS and LSA of response words to given cues for the 10 strongest responses in the free association norms.

	rank of response in free association									
k	1	2	3	4	5	6	7	8	9	10
	Word Association Space (WAS)									
10	86	187	213	249	279	291	318	348	334	337
50	13	36	49	62	82	98	106	113	125	132
100	6	17	26	36	43	62	65	73	78	85
200	3	8	15	20	28	39	40	48	56	58
300	2	6	12	16	21	31	35	38	43	49
400	1	5	10	14	19	27	32	35	38	44
	Latent Semantic Analysis (LSA)									
10	678	701	683	738	810	863	839	861	887	939
50	270	375	388	426	495	600	594	565	596	687
100	171	280	327	373	455	515	481	455	567	622
200	140	223	272	370	395	447	418	444	511	581
300	132	207	239	355	397	451	418	459	528	557

strongest associate to ‘baby’ in the free association norms, so ‘crib’ has rank 8 for the cue ‘baby’. Using the vectors from WAS, the rank of the similarity of a specific cue-response pair was computed by ranking the similarity among the similarities of the specific cue to all other possible responses. For example, the word ‘crib’ is the 2nd closest neighbor to ‘baby’ in WAS, so ‘crib’ has rank 2 for the cue ‘baby’. In this example, WAS has put ‘baby’ and ‘crib’ closer together than might be expected on the basis of free association norms. In Table 1, we compare the ranks from WAS to the ranks in the free association norms by computing the average of the ranks in WAS for the 10 strongest responses in the free association norms. The averaging was computed by the median to avoid excessive skewing of the average by a few high ranks. An additional variable that is tabulated in Table 1 is k, the number of dimensions of WAS.

There are three trends to be discerned in Table 1. First, it can be observed that for 400 dimensions, the strongest responses to the cues in free association norms are predominantly the closest neighbors to the cues in WAS. Second, responses that have higher ranks in free association have on average higher ranks in WAS. However, the output ranks in WAS are in many cases far higher than the output ranks in

Table 2
The five nearest neighbors in WAS for the first 40 cues in the Russell & Jenkins (1954) norms.

	neighbor				
Cue	1	2	3	4	5
Afraid	scare(1)[7]	fright(4)[14]	fear(2)[1]	scared[2]	ghost(5)[106]
Anger	mad(1)[1]	angry	rage(5)[4]	enrage	fury[21]
Baby	child(1)[2]	crib(8)[13]	infant(6)[7]	cradle	diaper(13)
bath	clean(2)[1]	soap(7)[3]	water(3)[2]	dirty[7]	suds[49]
beautiful	pretty(1)[2]	ugly(2)[1]	cute[39]	girl(4)	flowers[10]
bed	sleep(1)[1]	tired(11)[13]	nap	rest[5]	doze
bible	god(1)[1]	church(3)[3]	religion(4)[4]	Jesus(5)[8]	book(2)[2]
bitter	sweet(1)[1]	sour(2)[2]	candy	lemon(5)[7]	chocolate[4]
black	white(1)[1]	bleach	color(3)[7]	dark(2)[2]	minority
blossom	flower(1)[1]	petals[46]	rose(5)[7]	tulip	daisy
blue	color(5)[4]	red(3)[2]	jeans	crayon	pants
boy	girl(1)[1]	guy	man(4)[2]	woman	nephew[54]
bread	butter(1)[1]	toast[19]	rye[26]	loaf(3)[16]	margarine
butter	bread(1)[1]	toast(6)[18]	rye	peanut	margarine(2)[34]
butterfly	bug(15)[10]	insect(6)[2]	fly(4)[5]	roach[76]	beetle
cabbage	green(4)[7]	food(10)[4]	vegetable(2)[3]	salad(12)[5]	vegetables
carpet	floor(2)[2]	tile(15)	rug(1)[1]	ceiling	sweep[14]
chair	table(1)[1]	seat(4)[4]	sit(2)[2]	couch(3)[20]	recliner
cheese	cracker(2)	cheddar(6)[23]	Swiss(7)[19]	macaroni[39]	pizza
child	baby(1)[1]	kid(2)[7]	adult(3)[3]	young(8)[6]	parent(6)[11]
citizen	person(1)[3]	country(3)[5]	people[7]	flag[12]	American(2)[4]
city	town(1)[1]	state(2)[3]	country(9)[4]	New York(4)	Florida
cold	hot(1)[1]	ice(2)[5]	warm(6)[3]	chill	pepsi
comfort	chair(3)[1]	table	seat	couch(2)[26]	sleep[7]
command	tell(4)[7]	army(5)[2]	rules	navy[17]	ask[22]
cottage	house(1)[1]	home(4)[4]	cheese(2)[3]	cheddar	Swiss
dark	light(1)[1]	bulb	night(2)[2]	lamp	day
deep	water(3)[3]	ocean(2)[6]	faucet	pool[53]	splash
doctor	nurse(1)[1]	physician(5)[15]	surgeon(6)	medical[83]	stethoscope[21]
dream	sleep(1)[1]	fantasy(4)[19]	bed[7]	nap	tired[92]
eagle	bird(1)[1]	chirp	blue jay	nest(10)[5]	sparrow[30]
earth	planet(2)[8]	mars[14]	Jupiter[97]	Venus[50]	Uranus
eating	food(1)[1]	eat[30]	hungry(3)[4]	restaurant[75]	meal[30]
foot	shoe(1)[1]	sock[16]	toe(2)[3]	sneaker	leg(5)[4]
fruit	orange(2)[3]	apple(1)[1]	juice(9)[12]	citrus[35]	tangerine[55]
girl	boy(1)[1]	guy(6)	man[9]	woman(3)[2]	pretty(4)[6]
green	grass(1)[1]	lawn[41]	cucumber	vegetable[76]	spinach[76]
hammer	nail(1)[1]	tool(2)[7]	wrench	screwdriver	pliers[21]
hand	finger(1)[2]	arm(3)[3]	foot(2)[1]	leg(13)[11]	glove(4)[4]
hard	soft(1)[1]	easy(3)[3]	difficult[19]	difficulty	simple

Note: numbers in parentheses and square brackets indicate ranks of responses in norms of Nelson et al. (1998) and Russell & Jenkins (1954) respectively.

free association. For example, with 400 dimensions, the third largest response in free association is on average the 10th closest neighbor in WAS. Third, for smaller dimensionalities, the difference between the output order in free association and WAS becomes larger.

To summarize, given a sufficiently large number of dimensions, the strongest response in free association is represented (in most cases) as the closest neighbor in WAS. The other close neighbors in WAS are not necessarily associates in free association (at least not direct associates).

To get a better idea of the kinds of neighbors words have in WAS, in Table 2, we list the first five neighbors in WAS (using 400 dimensions) to 40 cue words. For all neighbors listed in the table, if they were associates in the free association norms of Nelson et al., then the corresponding rank in the norms is given between parentheses. Since all the 40 cue words are cue words used in the free association norms of Russell and Jenkins (1954), we also list the ranks in those norms between square brackets. The comparison between these two databases is interesting because Russell and Jenkins allowed participants to generate as many responses they wanted for each cue while the norms of Nelson et al. contain first responses only. We suspected that some close neighbors in WAS are not direct associates in the Nelson et al. norms but that they would have been valid associates if participants were allowed to give more than one association per cue. In Table 3, we list the percentages of neighbors in WAS of the 100 cues of the Russell and Jenkins norms (only 40 were shown in Table 2) that are valid/invalid associates according to the norms of Nelson et al. and/or the norms of Russell and Jenkins.

The last row shows that a third of the 5th closest neighbors in WAS are not associates according to the norms of Nelson et al. but that are associates according to the norms of Russell and Jenkins.

Table 3
Percentages of responses of WAS model that are valid/invalid in Russell & Jenkins (1954) and Nelson et al. (1998) norms

	neighbor				
word association norms	1	2	3	4	5
valid in Nelson et al.	96	73	61	45	33
valid in Jenkins et al.	96	83	79	69	64
valid in either Nelson et al. or Jenkins et al.	99	86	82	73	66
invalid in Nelson et al. but valid in Jenkins et al.	3	13	21	28	33

Therefore, some close neighbors in WAS are valid associates depending on what norms are consulted.

However, some close neighbors in WAS are not associates according to either norms. For example, ‘angry’ is the 2nd neighbor of ‘anger’ in WAS. These words are obviously related by word form but they do not to appear as associates in free association tasks because associations of the same word form tend to be edited out by participants. Because these words have similar associative structures, WAS puts them close together in the vector space.

Also, some close neighbors in WAS are not direct associates of each other but are indirectly associated through a chain of associates. For example, the pairs ‘blue-pants’ , ‘butter-rye’, ‘comfort-table’ are close neighbors in WAS but are not directly associated with each other. It is likely that because WAS is sensitive to the indirect relationships in the norms, these word pairs were put close together in WAS because of the indirect associative links through the words ‘jeans’, ‘bread’ and ‘chair’ respectively. In a similar way, ‘cottage’ and ‘cheddar’ are close neighbors in WAS because cottage is related (in one meaning of the word) to ‘cheese’, which is an associate of ‘cheddar’.

In Table 1, we also analyzed the correspondence between the similarities in the LSA space by Landauer and Dumais (1997) with the order of output in free association. As can be observed in the table, the rank of the response strength of the free association norms clearly has an effect on the ordering of similarities in LSA: strong associates are closer neighbors in LSA than weak associates. However, the overall correspondence between predicted output ranks in LSA and ranks in the norms is weak. The overall weaker correspondence between the norms and similarities for the LSA approach than the WAS approach highlights one obvious difference between the two approaches. Because WAS is based explicitly on free association norms, it is expected and shown here that words that are strong associates are placed close together in WAS whereas in LSA, words are placed in the semantic space in a way more independent from the norms.

Semantic/ Associative Similarity Relations

In the priming literature, several authors have tried to make a distinction between semantic and associative word relations in order to tease apart different sources of priming (e.g. Burgess & Lund, 2000; Chiarello, Burgess, Richards & Pollock, 1990; Shelton & Martin, 1992). Burgess and Lund (2000) have argued that the word association norms confound many types of word relations, among them, semantic and associative word relations. Chiarello et al. (1990) give “music” and “art” as examples of words that are semantically related because the words are rated to be members of the same semantic category (e.g. Battig & Montague, 1969). However, they claim these words are not associatively related because they are not direct associates of each other (according to the various norm databases that they used). The words “bread” and “mold” were given as examples of words that are not semantically related because they are not rated to be members of the same semantic category but only associatively related (since “bread” is an associate of “mold”). Finally, “cat” and “dog” were given as examples of words that are both semantically and associatively related.

We agree that the responses in free association norms can be related to the cues in many different ways, but it seems very hard and perhaps counterproductive to classify responses as purely semantic or purely associative[4]. For example, word pairs might not be directly but indirectly associated through a chain of associates. The question then becomes, how much semantic information do the free association norms contain beyond the direct associations? Since WAS is sensitive to the indirect associative relationships between words, we took the various examples of word pairs given by Chiarello et al. (1990) and Shelton and Martin (1992) and computed the WAS similarities between these words for different dimensionalities as shown in Table 4.

In Table 4, the interesting comparison is between the similarities for the semantic only related word pairs[5] (as listed by Chiarello et al., 1990) and 200 random word pairs. The random word pairs were selected to have zero forward and backward associative strength.

It can be observed that the semantic only related word pairs have higher similarity in WAS than the random word pairs. Therefore, even though Chiarello et al. (1990) have tried to create word pairs that were only semantically related, WAS can distinguish between these not directly associated word pairs and not directly associated random word pairs. This is because WAS is sensitive to indirect associative relationships between words. The Table also shows that for low dimensionalities, there is not as much difference between the similarity of word pairs that are semantically only and associatively only related. For higher dimensionalities, this difference becomes larger as WAS becomes more sensitive in representing more of the direct associative relationships.

To conclude, it is difficult to distinguish between pure semantic and pure associative relationships. What some researchers previously have considered to be pure semantic word relations, were word pairs that were related in their meaning but that were not directly associated with each other. This does not mean however that these words are not associatively related because the information in free association norms goes beyond that of direct associative strengths. In fact, the similarity structure of WAS turns out to be sensitive to the similarities that were argued by some researchers to be purely semantic.

Table 4
Average similarity between word pairs with different relations: semantic, associative, and semantic + associative

			k			
Relation	#pairs	B_{ij}[1]	10	50	200	400
Random	200	.000 (.000)	.340 (.277)	.075 (.178)	.024 (.064)	.017 (.048)
			Word pairs from Chiarello et al. (1990)			
Semantic only	33	.000 (.000)	.730 (.255)	.457 (.315)	.268 (.297)	.215 (.321)
Associative only	43	.169 (.153)	.902 (.127)	.830 (.178)	.712 (.262)	.666 (.289)
Semantic + Associative	44	.290 (.198)	.962 (.053)	.926 (.097)	.879 (.180)	.829 (.209)
			Word pairs from Shelton and Martin (1992)			
Semantic only	26	.000 (.000)	.724 (.235)	.448 (.311)	.245 (.291)	.166 (.281)
Semantic + Associative	35	.367 (.250)	.926 (.088)	.929 (.155)	.874 (.204)	.836 (.227)

Note: standard deviations given between parentheses
1: B_{ij} = average forward and backward associative strength = (A_{ij} + A_{ji}) / 2

Table 5
Average Between and Within Category Similarities in WAS of Murdock's (1976) Semantic Categories

normalize	Between	Within	ratio (between/within)
N	.0003 (.0012)	.0061 (.0471)	17.8
Y	.0182 (.0107)	.3418 (.3459)	18.8

Note: standard deviations between parentheses

Capturing Between/Within Semantic Category Differences in WAS

In this section, we give an additional demonstration that the space formed by WAS is sensitive to semantic information. Murdock's (1976) collected 32 semantic categories with each 32 category members. Examples of categories are 'body parts', 'ships', 'birds', 'fruits', and 'tools'. Members of the first category were for example 'leg', 'arms', 'head', 'eye' and members of the second category were for example 'sailboat', 'destroyer', 'battleship'. If WAS is sensitive to the categorical structure of these semantic norms, then the within category similarity should on average be higher than the between category similarity. Similarity was computed by the inner product between word vectors. The within category similarity was calculated by averaging the similarities of all possible word pairs within a category. Similarly, the between category similarity was calculated by averaging the similarities of all possible word pairs that fell in different categories. In Table 5, the between and within category similarities are shown. Note that the within category similarity is 18 times higher than the between category similarity suggesting that the similarity structure of WAS is well suited to represent semantic categorical information. The row labeled 'not normalized' refers to the space used in part I of the research where the vector lengths are not normalized. In the second row, the table shows that when the vector lengths are normalized, the ratio of within to between category similarity is equally high. This observation becomes important in part II of this research, where we do normalize the vector lengths.

Predicting Memory Performance

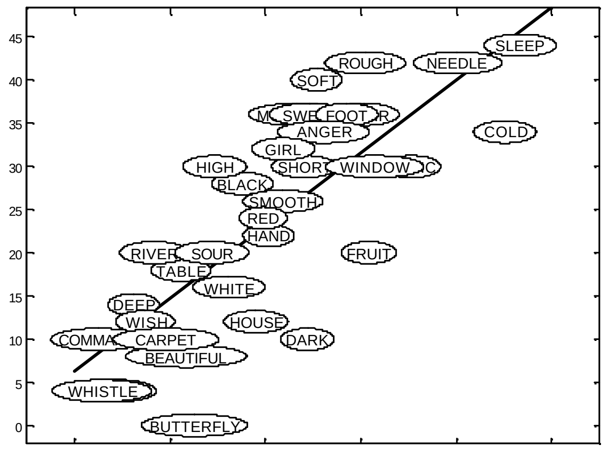
Predicting Results from Deese

In a classic study by Deese (1959b), the goal was to predict the intrusion rates of words in free recall. Participants studied the 15 strongest associates to each of 36 critical lures while the critical lures themselves were not studied. In a free recall test, some critical lures (e.g. "sleep") were falsely recalled about 40% of the time while other critical lures (e.g. "butterfly") were never falsely recalled. Deese was able to predict the intrusion rates for the critical lures on the basis of the average associative strength from the studied associates to the critical lures and obtained a correlation of R=0.8. Since Deese could predict intrusion rates with word association norms, it was expected that that the WAS vector space derived from the association norms could also predict intrusion rates. The idea here is that critical items that are closely related to list words are more likely to appear as intrusions in free recall than critical items that are not closely related to list words. The average similarity was computed between each critical lure vector and list word vectors using different dimensionalities. In Figure 3, a scatter plot shows the relationship between the similarity and the intrusion rates as observed by Deese (here, the number of dimensions was 400). The obtained correlation was R=0.775. In Table 6, the correlations for other dimensionalities are listed. The correlation

Table 6
Correlations between the average similarity of critical and list words and the intrusion rates observed by Deese (1959b)

k	WAS	LSA
10	.386	.210
50	.519**	.189
100	.617**	.154
200	.691**	.204
300	.682**	.174
400	.775**	-

Note: ** Correlation is significant at the 0.01 level (2-tailed)



3. The average similarity between critical item and list item in WAS can predict the intrusion rates for the critical item as observed by Deese (1959b).

decreases with decreasing number of dimensions. This might happen because a smaller dimensional space has less room to place 5000 words so that the resulting similarity structure does not capture as well the differences in observed intrusion rates. The table also shows the correlations when the vectors are taken from LSA. It can be seen that similarity structure of LSA does not correlate as well with the intrusion rates as WAS. Also, the effect of varying the number of dimensions does not seem to affect the correlations.

Predicting Extralist Cued Recall

In extralist cued recall experiments, after studying a list of words, subjects are presented with cues that can be used to retrieve words from the study list. The cues themselves are novel words that were not presented during study and they typically are associatively related to one of the studied words. The degree to which a cue is successful in retrieving a particular target word is a measure of interest because this might be related to the associative/semantic overlap between cues and their targets. Research in this paradigm (e.g. Nelson & Schreiber, 1992; Nelson, Schreiber, & McEvoy, 1992; Nelson, McKinney, Gee, & Janczura, 1998; Nelson & Xu, 1995) has already shown that the associative strength between cue and target is one important predictor for the percentage correctly recalled targets. Therefore, we expect that the WAS similarity between cues and targets are correlated to the percentages of correct recall in these experiments. We used a database containing the percentages correct recall for 1115 cue-target pairs from over 29 extralist cued recall experiments from Doug Nelson's laboratory (Nelson & Zhang, submitted; Nelson, personal communication). The correlations between the WAS similarity and observed recall rates for different dimensionalities are shown in Table 7.

The best result was a small but significant correlation of .36 using 400 dimensions. The correlations decreased with decreasing number of dimensions. Since a smaller number of dimensions limits the ways in which 5000 words can be placed in the space, it is possible that this factor explains the limiting effect on the correlation. The table also shows the correlations when vectors from the LSA space were taken. The correlations with the LSA vectors were less high than with WAS but were relatively close in value at 300 dimensions. This suggests that both WAS and LSA can be used as part of a process model to predict cued recall results.

Table 7
Correlations between the similarity of cued recall word pairs and percentage correct recall rates using WAS and LSA

k	WAS	LSA
10	.051	.004
50	.214**	.119**
100	.274**	.167**
200	.335**	.220**
300	.342**	.252**
400	.360**	-

Note: ** Correlation is significant at the 0.01 level (2-tailed)

Discussion

By a statistical analysis of a large database of free association norms, the Word Association Space (WAS) was developed. In this space, words that have similar associative structures are placed in similar regions of the space. We showed that the output order of words in free association norms is preserved to some degree in WAS: first associates in the norms are likely to be close neighbors in WAS. There are some interesting differences between the similarity structure of WAS and the associative strengths of the words in the norms. Words that are not directly associated can be close neighbors in WAS when the words are indirectly associatively related through a chain of associates. Also, in some cases, words that are directly associated in the norms are not close neighbors in WAS at all (although these are exceptions). This makes WAS not a good model for the task of predicting free association data. However, it is important to realize that WAS was not developed as a model of free association (e.g. Nelson & McEvoy, Dennis, in press) but rather as a model based on free association.

The WAS approach is an additional method available to place words in a psychological space. It differs from the LSA and HAL approaches in several ways. LSA and HAL are automatic methods and do not require any extensive data collection of ratings or free associations. With LSA and HAL, tens of thousands of words can be placed in the space, whereas in WAS, the number of words that can be placed depends on the number of words that can be normed. It took Nelson et al. (1998) more than a decade to collect the norms, highlighting the enormous human overhead of the method.

Another difference is that LSA and HAL have the potential to model the learning process a language learner goes through. For example, by feeding the LSA or HAL model successively larger chunks of text, it can be simulated what the effect learning has on the similarity structures of words in LSA or HAL. In WAS, it is in principle possible to model a language learning process by collecting free association norms for participants at different stages of the learning process. In practice however, such an approach would not easily be accomplished.

We think that the WAS, LSA, and HAL approaches to creating semantic spaces are all useful for theoretical and empirical research. It might be that the usefulness of a particular space will depend on the task it is applied to. Since the free association norms have been an integral part in predicting episodic memory phenomena (e.g. Cramer, 1968; Deese, 1965; Nelson, Schreiber, & McEvoy, 1992), it was assumed that a vector space based on free association norms would be an especially useful construct to model memory phenomena. In this research, we have already shown with simple geometric operations how the similarity structure of WAS can predict to some degree the intrusion rates observed by Deese (1959b) in his classic false memory experiment as well as the percentages of correct recall in cued recall experiments. This suggests to us that WAS forms a useful representational basis for memory models that are designed to store and retrieve words as vectors of feature values. In part II of this research, we will combine the semantic space of WAS with a process model for recognition memory. This will allow us to model the processes of recognition memory and gives us a principled way to represent words by vectors. The assumption of representing words by vectors in memory models is relatively old. However, in most memory modeling, the vectors representing words are arbitrarily chosen and are not based on or derived by some analysis of the meaning of actual words in our language. In part II, it is expected that a memory model based on these semantic vectors from WAS will be useful to make predictions about the effects of varying semantic similarity in memory experiments.

Appendix

Let the matrix A represent the information from the free association norms with A_{ij} representing the relative frequency with which participants generate response j with cue i. The idea is to use the information in the matrix of the free association norms to place the n words in a high dimensional space by applying singular value decomposition. We first transformed A to a new matrix T by symmetrizing A and by adding the two-step indirect associative strengths[6] from the cue to response and from response to cue:

$$T_{ij} = A_{ij} + A_{ji} + \sum_k A_{ik} A_{kj} + \sum_k A_{jk} A_{ki} \qquad (1)$$

The matrix T is symmetric: $T_{ij} = T_{ji}$. It is possible to decompose any square symmetric matrix T into a product of three matrices by using a special case of the singular value decomposition method[7]:

$$T = U_0 D_0 U_0{}' \qquad (2)$$

Here, U'_0 denotes the transpose of U_0. When the matrix T has size n x n (i.e., n rows and n columns), then U_0 and D_0 are also size n x n. The columns of matrix U_0 are orthonormal and contain the N eigenvectors. The matrix D_0 is diagonal and contains the n singular values. It is customary to let the first diagonal entry contain the largest eigenvalue followed by eigenvalues in decreasing order.

The purpose of this linear decomposition is to approximate matrix T by matrices with a much smaller number of singular values and singular vectors:

$$\hat{T} = UDU' \qquad (3)$$

Here, D is the k x k diagonal matrix containing only the k largest (k << n) singular values of D_0. U is the n x k matrix that contains only the first k eigenvector columns of U_0. We represent words by the column vectors of the matrix X, which is formed by weighting the eigenvectors with the eigenvalues:

$$X = UD \qquad (4)$$

The matrix X represents the high dimensional vector space that is called 'Word Association Space'. Each column vector of X represents the location of a word in the space.

Notes

[1] The fact that HAL uses a much smaller window in which to calculate co-occurrence statistics than in LSA might explain the finding that HAL is more sensitive to the grammatical aspects of meaning: nouns, prepositions and verbs cluster together in the contextual space of HAL.

[2] The number is dimensions that can be extracted is constrained by various computational aspects. We were able to extract only the first 400 dimensions for WAS.

[3] The correlation between the log Kucera and Francis frequency and the log of the number of times a word was produced in the free association norms was 0.53.

[4] Since responses in word association tasks are by definition all associatively related to the cue, it is not clear how it is possible to separate the responses as semantically and associatively related.

[5] Some word pairs in the semantic only conditions that were not directly associated according to various databases of free association norms were actually directly associated using the Nelson et al. (1998) database. These word pairs were excluded from the analysis.

[6] We have added the indirect associations to the word association matrix because we have found that this leads to vector spaces that better preserve the order of associative strengths of the original word association matrix. At this time, it is not clear what the reason is for the advantage of adding the indirect strengths. More research is needed to investigate the influence of this preprocessing step on the similarity structure of the resulting vector space.

[7] the SVD method is more general and can decompose any rectangular or asymmetric matrix. For a discussion showing the relationship between SVD and relationship to multidimensional scaling see Bartell, Cottrell, and Belew (1992).

References

Anisfeld, M., & Knapp, M. (1968). Association, synonymity, and directionality in false recognition. Journal of Experimental Psychology, 77, 171-179.

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 Monograph, 80(3), 1-46.

Bartell, Brian B., Cottrell, G.W. & Belew, R. (1992) Latent Semantic Indexing is an Optimal Special Case of Multidimensional Scaling. In Proceedings of Special Interest Group on Information Retrieval, Copen-hagen, Denmark, ACM Press.

Bousfield, W.A. (1953). The occurrence of clustering in the recall of randomly arranged associates. Journal of General Psychology, 49, 229-240.

Bower, G.H. (1967). A multicomponent theory of the memory trace. In K.W. Spence & J.T. Spence (Eds.), The psychology of learning and motivation, Vol 1. New York: Academic Press.

Burgess, C., Livesay, K., and Lund, K. (1998). Explorations in context space: Words, sentences, discourse. Discourse Processes, 25, 211-257.

Burgess, C., & Lund, K. (2000). The dynamics of meaning in memory. In E. Dietrich and A.B. Markman (Eds.), Cognitive dynamics: conceptual and representational change in humans and machines. Lawrence Erlbaum.

Chiarello, C., Burgess, C., Richards, L., & Pollock, A. (1990). Semantic and associative priming in the cerebral hemispheres: Some words do, some words don't, …sometimes, some places. Brain and Language, 38, 75-104.

Canas, J. J. (1990). Associative strength effects in the lexical decision task. The Quarterly Journal of Experimental Psychology, 42, 121-145.

Caramazza, A., Hersch, H., & Torgerson, W.S. (1976). Subjective structures and operations in semantic memory. Journal of verbal learning and verbal behavior, 15, 103-117.

Cramer, P. (1968).Word Association. NY: Academic Press.

Deese, J. (1959a). Influence of inter-item associative strength upon immediate free recall. Psychological Reports, 5, 305-312.

Deese, J. (1959b). On the prediction of occurrences of particular verbal intrusions in immediate recall. Journal of Experimental Psychology, 58, 17-22.

Deese, J. (1960). Frequency of usage and number of words in recall: the role of association. Psychological Reports, 7, 337-344.

Deese, J. (1962). On the structure of associative meaning. Psychological Review, 69, 161-175.

Deese, J. (1965). The structure of associations in language and thought. Baltimore, MD: The Johns Hopkins Press.

Derweester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41, 391-407.

Eich, J.M. (1982). A composite holographic associative recall model. Psychological Review, 89, 627-661.

Jenkins, J.J., Mink, W.D., & Russell, W.A. (1958). Associative clustering as a function of verbal association strength. Psychological Reports, 4, 127-136.

Herriot, P. (1974). Attributes of memory. London: Methuen.

Hintzman, D.L. (1984). Minerva 2: a simulation model of human memory. Behavior Research Methods, Instruments, and Computers, 16, 96-101.

Krumhansl, C.L. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. Psychological Review, 85, 445, 463.

Kucera, H., & Francis, W.N. (1967). Computational analysis of present-day American English. Providence, RI: Brown University Press.

Landauer, T.K., & Dumais, S.T. (1997). A solution to Plato’s problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.

Landauer, T.K., Foltz, P., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25, 259-284.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instruments, and Computers, 28, 203-208.

Morton, J.A. (1970). A functional model for memory. In D.A. Norman (Ed.), Models of human memory. New York: Academic Press.

Murdock, B.B. (1976). Item and order information in short-term serial memory. Journal of Experimental Psychology: General, 105, 191-216.

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

Neely, J.H. (1991). Semantic priming effects in visual word recognition: a selective review of current findings and theories. In D. Besner & G.W. Humphreys (Eds.), Basic processes in reading: Visual word recognition (pp. 264-336). Hillsdale, NJ: Lawrence Erlbaum Associates.

Nelson, D.L., Bennett, D.J., & Leibert, T.W. (1997). One step is not enough: making better use of association norms to predict cued recall. Memory & Cognition, 25, 785-706.

Nelson, D.L., McEvoy, C.L., & Dennis, S. (in press), What is and what does free association measure? Memory & Cognition.

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

Nelson, D.L., McKinney, V.M., Gee, N.R., & Janczura, G.A. (1998). Interpreting the influence of implicitly activated memories on recall and recognition. Psychological Review, 105, 299-324.

Nelson, D.L., & Schreiber, T.A. (1992). Word concreteness and word structure as independent determinants of recall. Journal of Memory and Language, 31, 237-260.

Nelson, D.L., Schreiber, T.A., & McEvoy, C.L. (1992). Processing implicit and explicit representations. Psychological Review, 99, 322-348.

Nelson, D.L., Xu, J. (1995). Effects of implicit memory on explicit recall: Set size and word frequency effects. Psychological Research, 57, 203-214.

Nelson, D.L., & Zhang, N. (submitted). The ties that bind what is known to the recall of what is new.

Norman, D.A., & Rumelhart, D.E. (1970). A system for perception and memory. In D.A. Norman (Ed.), Models of human memory. New York: Academic Press.

Osgood, C.E., Suci, G.J., & Tannenbaum, P.H. (1957). The measurement of meaning. Urbana: University of Illinois Press.

Palermo, D.S., & Jenkins, J.J. (1964). Word association norms grade school through college. Minneapolis: University of Minnesota Press.

Pike, R. (1984). Comparison of convolution and matrix distributed memory systems for associative recall and recognition. Psychological Review, 91, 281-293.

Postman, L. (1975). Verbal learning and memory. Annual Review of Psychology, 26, 291-335.

Rips, L.J., Shoben, E.J., & Smith, E.E. (1973). Semantic distance and the verification of semantic relations. Journal of verbal learning and verbal behavior, 12, 1-20.

Romney, A.K., Brewer, D.D., & Batchelder, W.H. (1993). Predicting clustering from semantic structure. Psychological Science, 4, 28-34.

Russell, W.A., & Jenkins, J.J. (1954). The complete Minnesota norms for responses to 100 words from the Kent-Rosanoff word association test. Tech. Rep. No. 11, Contract NS-ONR-66216, Office of Naval Research and University of Minnesota.

Schwartz, R.M., & Humphreys, M.S. (1973). Similarity judgments and free recall of unrelated words. Journal of Experimental Psychology, 101, 10-15.

Shelton, J.R., & Martin, R.C. (1992). How semantic is automatic semantic priming? Journal of Experimental Psychology: Learning, Memory, and, Cognition, 18, 1191-1210.

Shiffrin, R.M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. Psychonomic Bulletin & Review, 4, 145-166.

Underwood, B.J. (1965). False recognition produced by implicit verbal responses. Journal of Experimental Psychology, 70, 122-129.

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

Wickens, D.D. (1972). Characteristics of word encoding. In A.W. Melton & E. Martin (Eds.), Coding processes in human memory. Washington, D.C.: V.H. Winston, pp. 191-215.

Part II: Predicting Memory Performance with Word Association Spaces

Many memory models assume that the semantic and physical features of words can be represented by collections of features abstractly represented by vectors (e.g. Eich, 1982; Murdock, 1982; Pike, 1984; Hintzman, 1988; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997, 1998). Most of these vector memory models are process oriented; they explicate the processes that operate on memory representations without explicating the origin of the representations themselves: the different attributes of words are typically represented by random vectors that have no formal relationship to the words in our language. The first goal of this research was to develop vector representations that capture the aspects of the meaning of words and vector representations that capture the physical aspects of words such as orthography and/or phonology. As opposed to the vector representations used by many memory models, the semantic and physical features in these representations do have formal relationships to words in the English language. The second goal of this research was to combine these representations with a process model for memory. This part of the research was built on previous research with the REM model (Shiffrin & Steyvers, 1997, 1998) in which a framework was laid out for a process model of episodic memory. With this processing model, we aimed to provide a qualitative account for various recognition memory phenomena found in the literature, as well as the results of the experiments reported in this paper. In addition to the physical and semantic attributes, word frequency was a factor that had to be taken into account in the modeling and experiments, because word frequency variation produces large effects on recognition memory performance. In summary, we aim to provide qualitative accounts for differences in individual word performance in recognition memory based on semantic features, physical features, and the natural language frequency of the words that are studied and tested.

Semantic and Physical Similarity Effects in Memory

One way to investigate the role of semantic features involves varying the semantic similarity between study and test words, often carried out within the 'false memory paradigm'. Following the classic experiments by Deese (1959a, b), Roediger and McDermott (1995) revived interest in this paradigm (e.g. Brainerd, & Reyna, 1998, 1999; Payne, Elie, Blackwell, & Neuschatz, 1996; Schacter, Verfaellie, & Pradere, 1996; Tussing & Green, 1997). In the typical false memory experiment, participants study words that are all associatively and/or semantically related to a non-studied critical word. In a subsequent recognition test, the critical word typically lead to a higher false alarm rate than that for unrelated foils (and sometimes quite high in comparison to that for studied words). In a free recall test, participants falsely intrude the critical word at a rate higher than unrelated words (and sometimes at rates approaching those for studied words). These studies show that memory errors can be strongly influenced by semantic similarity.

Phonetic and orthographic similarity has been shown to play a role in free recall (Watkins, Watkins, & Crowder, 1974; Brown & McNeill, 1966) and cued recall (Bregman, 1968; Laurence, 1970; Nelson & Brooks, 1973; Wickens, Ory, & Graf; 1970). In recognition memory, acoustically/orthographically similar distractors lead to higher false alarm rates than acoustically/orthographically dissimilar distractors (Buschke & Lenon, 1969; Cermak, Schnorr, Buschke & Atkinson, 1970; Davies & Cubbage, 1976; Runquist & Blackmore, 1973). These studies show that memory errors can be based on similarity of orthographic, phonological, and semantic features of words, and emphasizes the need to include mechanisms reflecting these factors in memory models.

We now discuss four of the many explanations for semantic and orthographic/ phonological similarity effects in memory; these explanations are not mutually exclusive:

Generation of episodic traces at study. Underwood (1965) proposed that during study of words, participants generate "implicit associative responses" (IAR's) which might be stored as episodic traces in memory. If the study list contains many fruit words (e.g. "apple", "pear", "banana" etc.) but not the word "fruit" itself, the word "fruit" might be so strongly evoked in mind by all the fruit words that the word "fruit" might be actually stored in memory as if it had been presented during study. This essentially locates the false memory effect at storage. Little detail has as yet been provided for the underlying mechanism of IAR's. There is some evidence that a strong version of this mechanism is not sufficient to explain false memory effects: If it is assumed that the fruit study list always leads to storage of the word "fruit" in memory, then testing "fruit" as a distractor should lead to the same level of familiarity as testing "fruit" as a target when the word was actually presented on the study list. Miller and Wolford (1999) found that participants can distinguish between critical words tested as distractors and critical words tested as targets, thus casting doubt on the strong version of the IAR theory. However, these results are compatible with a mechanism in which it is assumed that IAR's lead to weaker traces in memory than actually presented items.

Shiffrin, Huber, and Marinelli (1995) varied the category size of studied words; categories either contained semantically similar words or orthographically similar words. They found that false recognitions for both semantically and orthographically similar distractors increased as category size increased, and argued that it was unlikely these category length effects were due to IAR's. First, the category words were spaced throughout a very long study list, making it difficult for participants to perceive the underlying categories. Participants reported that they were not aware of the underlying category structures, in almost all instances. Second, it is probably less likely that the IAR mechanism would apply in explaining false memory effects based on physical similarity, because most explicit or conscious coding in memory studies appears to be based on semantic content. For example, when the study list contains "BEG", "BOG", "BIG", and "BUG" spaced 20 or more items apart in a long list, it is rather unlikely that an elevated false alarm rate for "BAG" is due to participants explicitly thinking about the word "BAG" during study (although such phonological productions might well occur in massed study situations).

Based on such results, it seems likely that the IAR mechanism plays a significant role especially when similar study words are grouped together. When the IAR mechanism operates, and produces a memory trace for a word, such a trace would probably not be as strong as that produced by that same word actually presented.

Storage in lexical/semantic traces. the result of study of a category of related items might include not only storage of an explicit, episodic trace for the non-studied IAR word, but also storage in the lexical/semantic trace for that word. For example, the REM model for implicit memory (Schooler, Shiffrin, & Raaijmakers, in press) posits storage of context information in a word's lexical/semantic trace following its study; this could occur as well after IAR generation. For example, during study of many fruit words, the lexical entry for "fruit" (not presented during study) might be activated and might gain a small number of current context features. These context features represent the immediate situation and task. When the word "fruit" is tested, a false alarm might be generated because the current context matches the context features stored in the lexical trace for "fruit". Sommers and Lewis (1999) propose an account for phonological false memory effects that is similar to this notion of implicit activation. Neighboring words in phonological space gain activation from presentation of a study word. This was implemented with the NAM model (Luce & Pisoni, 1998). For example, studying the words "BEG", "BOG", "BIG", and "BUG" leads to enhanced activation of the words "BAG" in some phonological space. The idea is that because a word such as "BAG" has extra activation, the false alarm rate of this word (when tested as a distractor), will be increased relative to other words.

Storage of gist. Brainerd and Reyna (1998; 1999) have proposed in their Fuzzy trace theory that the presentation of study words leads to the storage of two kinds of traces in memory: verbatim and gist traces. Verbatim traces relate to the surface features (e.g. orthography, phonology) of individual words while gist traces relate more to the collective meaning of the studied material (Bransford & Franks, 1972). For example, studying words like "pillow", "dream", "bed", "snore" might lead to verbatim traces for each of these individual words and also a gist trace that could be interpreted as "sleep". Therefore, testing "sleep" as a distractor leads to high false alarms because it matches the stored gist. The focus of this theory has been to show the independent effects of the processes operating on the verbatim and gist traces. To date, the fuzzy trace theory has been implemented as a measurement model (see Brainerd, Reyna, & Mojardin, 1999), and not as a process model: the theory does not specify how gist and surface traces are extracted, stored and retrieved at test.

Global familiarity operating at retrieval. In global familiarity models such SAM (e.g. Gillund & Shiffrin, 1984), MINERVA (Hintzman, 1988) and REM (Shiffrin & Steyvers, 1997), it is assumed that study leads to separate traces in memory for every word presented. At retrieval, the stored traces are activated in proportion to their similarity to a test word, and the summed activations are used to make a recognition decision. In the REM instantiation, for example, words are represented by vectors of feature values that are assumed to contain among other attributes, phonological, orthographic and semantic features. The episodic traces that are stored in memory contain error-prone and/or incomplete copies of the features of the word vectors. The recognition process is based on a comparison of the probe to every trace in memory: a match value is calculated for each probe/trace comparison. The recognition decision is based on a function of the sum of these individual match values. A decision "old" is made when the sum exceeds a certain criterion, otherwise a decision “new” is made. An incorrect “old” recognition for a distractor can be expected when the probe features will match the features of several traces to such a degree that the sum of the match values exceeds the criterion. The global familiarity mechanism therefore explains the false memory effect as a retrieval effect.

Word frequency effects in recognition memory

Word frequency can be defined by counting the number of times a word occurs in samples of written text (Kucera and Francis, 1967). The number of times a word is experienced pre-experimentally, and/or the relative number of times a word is experienced pre-experimentally, have a large effect on memory performance even though experimental frequency and other factors are held constant. Low frequency words are better recognized than high frequency words (Glanzer & Bowles, 1976; Gorman, 1961; Kinsbourne & George, 1974; McCormack & Swenson, 1972; Shepard, 1976; Schulman & Lovelace, 1970). In addition, the hits (responding 'old' to a target) and false alarms (responding 'old' to a foil) typically exhibit a mirror effect: hits are higher for low than high frequency words, and false alarms are higher for high than low frequency words (e.g. McCormack & Swenson, 1972; Glanzer & Adams, 1990).

Word frequency is correlated with many other measures defined for words such as feature frequency, concreteness, the number of different meanings, recency, and the number of contexts in which they appear. Not surprisingly, then, quite a few mechanisms have been proposed to explain word frequency effects. We next discuss three of these:

Trace strength differences. One explanation for the word frequency effect is based on the strength of encoding. Mandler (1980) proposed that low frequency words are rehearsed more than high frequency words so that they are encoded better in memory. In a similar account, Glanzer and colleagues (Glanzer & Adams, 1990; Kim & Glanzer 1993) proposed that low frequency words attract more attention so that they are better encoded. This explanation (and others as well) does not explain why lists of high frequency words are free-recalled better than lists of low-frequency words (e.g., Gregg, 1976). However, in the SAM and REM models, recall operates not through a process of global activation (which applies to recognition) but instead through a search process involving steps of sampling and recovery. In these theories, recovery is superior for high frequency words, overcoming any other advantage that may favor low frequency words.

Feature frequency differences. An explanation for word frequency based on both coding and retrieval is based on feature frequency differences. This idea was explored in Shiffrin and Steyvers (1997). Landauer and Streeter (1973) showed that high and low frequency words are structurally different: on average, different features make up high and low frequency words. In Shiffrin and Steyvers (1997), the assumption was made that high frequency words tended to contain high frequency features, justified by the argument that high frequency words are encountered more often, hence insuring that their features are also encountered more often. In the REM model, the feature values for high frequency words were made more common than the feature values for low frequency words. Since a match of a rare feature in the probe and a trace was more diagnostic than a match of a common feature, the system predicted advantages for low frequency words (in recognition memory). In part III of this research, we will provide empirical support for this explanation by independently varying word frequency and feature frequency. To preview the results: words with equal word frequency are better remembered when the words consist primarily of low than high frequency features, a result consistent with the feature frequency hypothesis for word frequency effects.

Context differences. Since high frequency words occur more often than low frequency words, on average they also occur more recently than low frequency words (e.g. Scarborough, Cortese, & Scarborough, 1977). This can lead to more confusion in recognition memory for high frequency than low frequency words. That is, for high frequency words a large value of familiarity could arise correctly for targets, but incorrectly for foils due to a pre-experimentally recent occurrence. High frequency words also occur in a greater variety of contexts (Dennis, 1995) than low frequency words. In a model by Dennis and Humphreys (1998; submitted), this difference in context noise was used to predict word frequency effects.

It is entirely possible that all three of these word frequency accounts are valid (along with others we have not discussed) and that multiple mechanisms are operating simultaneously. The focus in this article will be word frequency effects due to feature frequency effects and context differences.

A memory model for semantic and orthographic similarity effects

The memory model in this research is based on the REM model that in its first inception was fit qualitatively to various basic recognition memory phenomena (Shiffrin & Steyvers, 1997, 1998). Later, Diller, Nobel, and Shiffrin (in press) fitted the model quantitatively to recognition and cued recall experiments. In more recent work, the model has been extended to various implicit memory tasks (e.g. Schooler, Shiffrin, & Raaijmakers, in press) and short-term priming (Huber, Shiffrin, Lyle, Ruijs, in press).

In the previous sections, it was established that both semantic and physical similarity between probe and memory traces are important determinants of memory performance: both semantically and physically similar distractor probes tend to produce higher false alarm rates than unrelated control words. In the three experiments in this paper, the role of semantic similarity, physical similarity and word frequency in recognition memory are investigated.

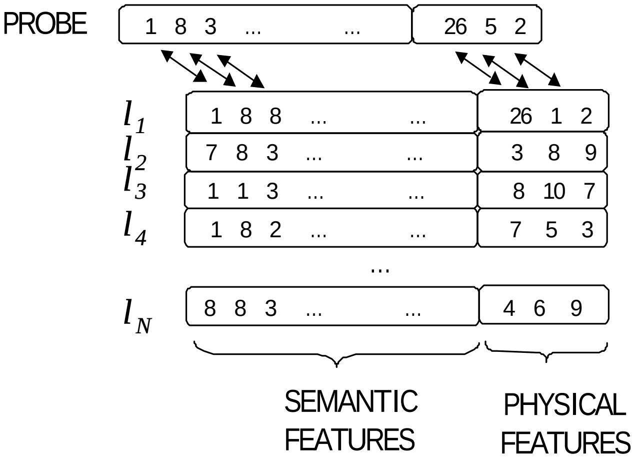

PROBE
1 8 3 ... ... 26 5 2
l_1 1 8 8 ... ... 26 1 2
l_2 7 8 3 ... ... 3 8 9
l_3 1 1 3 ... ... 8 10 7
l_4 1 8 2 ... ... 7 5 3
...
l_N 8 8 3 ... ... 4 6 9
SEMANTIC FEATURES
PHYSICAL FEATURES

Figure 1. Illustration of the memory model. The semantic and physical features of the probe are compared in parallel to corresponding features in all episodic traces in memory. The model calculates a likelihood ratio for each probe-trace comparison, expressing the match between probe and trace. The overall familiarity that forms the basis for recognition judgments is calculated by the sum of likelihood ratio's.

We have two goals: 1) using a version of the REM model, we hope to fit qualitatively the results from the three experiments reported in this paper. 2) we shall investigate the degree to which it is possible to predict differences in performance for individual words as opposed to groups of words. Because we have a process model operating on a representation of the semantic and physical attributes of words that is based on an analysis of actual words, we can make a priori predictions for individual words. This approach differs from that in which similarity constraints are imposed on arbitrary feature vectors.

Overview of Model

REM uses Bayesian principles to model the decision process in recognition memory. Words are stored in memory as episodic traces represented by vectors of feature values. We adopt the REM assumption that all information related to the study episode is stored in one trace; in this research, such information is defined to consist of semantic and physical features. At study, the presented word contacts its lexical/semantic trace, and an attempt is made to store the combination of the physical features and the features recovered from the lexical trace. The resultant episodic trace is an incomplete and error prone copy of these feature values. Retrieval operates by comparing in parallel the semantic and physical features of the test word to all traces, and measuring the featural overlap for each trace as illustrated in Figure 1.

The featural overlap for each trace contributes evidence to a likelihood ratio for each trace. In Shiffrin and Steyvers (1997), it was shown that the odds for 'old' over 'new' equaled the sum of the likelihood ratios divided by the number of traces involved in comparisons.

Two memory judgments

We borrow the procedure used by Brainerd and Reyna (1998) in which participants were instructed to give one of two memory judgments: standard recognition instructions and joint recognition instructions. With standard recognition instructions, participants were instructed to respond "yes" to targets and "no" to all distractors. With joint recognition instructions, participants were instructed to respond "yes" to targets and "yes" to all distractors that are related in meaning to one of the various themes of the words on the study list. They only had to respond "no" to unrelated distractors. We will refer to the two memory judgments that are generated under the standard recognition and joint recognition instructions as recognition and similarity judgments respectively.

Comparison of the results for recognition and similarity judgments allows investigation of the interplay between semantic and physical features, especially if one assumes that similarity judgments are based only on the matching of semantic information, and not physical information (as the instructions imply). We can test this assumption by modeling the similarity judgments with semantic features only, and modeling the recognition judgments with both semantic and physical features. Based on these assumptions, the difference between the recognition and similarity ratings measures the degree of reliance on physical features.

Semantic features

In part I, we showed how a semantic space was constructed by analyzing the statistical structure of word association norms. We borrowed the singular value decomposition technique (SVD) of the latent semantic analysis approach (LSA, Landauer and Dumais, 1997) to place words in a high dimensional semantic space. In LSA, semantic spaces are created by analyzing co-occurrence statistics of words appearing in different contexts in large text documents such as encyclopedia. The idea is that words similar in meaning appear in similar contexts (where context is defined as segments of connected text such as individual encyclopedic entries).

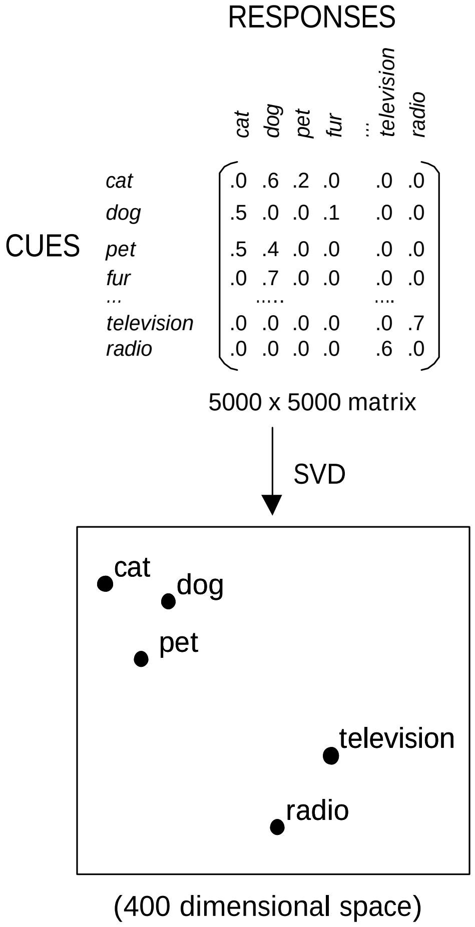


Figure 2. Illustration of the Word Association Space (WAS) approach. The singular value decomposition (SVD) method is applied on a large database of free association norms to place words in a high dimensional space. Words are placed in similar locations in a high dimensional psychological space when the associative relationships between words are similar.

In our approach, the SVD procedure was applied to the matrix of free associations for over 5000 words collected by Nelson, McEvoy, and Schreiber (1998). The result is that words that have similar associative structures are placed as points in similar regions of a 400 dimensional space as illustrated in Figure 2. To put it differently, each word was represented as a vector of 400 feature values with associatively similar words having similar feature values. Because the space was developed on word association norms, the space was named Word Association Space (WAS).

The basic distinction between LSA and WAS is that in the former approach, it was assumed that similar words occur in similar contexts, while in the latter approach, it was assumed that similar words have similar associative structures. Both conceptual frameworks are useful in empirical and theoretical research. The WAS approach was developed with the specific purpose of modeling memory phenomena. Since it has been established that the associative structure can predict recall (e.g. Cramer, 1968; Deese, 1959a,b, 1965), cued recall (e.g., Nelson, Schreiber, & McEvoy, 1992), and priming (Canas, 1990), we expected that the word association space formed by analyzing the free association norms is particularly useful to predict memory performance.

As described in part I, **WAS** is not a metric space in which distance measures dissimilarity. The SVD analysis that produced **WAS** is based on the idea that inner products represent similarity. Thus high frequency words, which are more similar to each other (as measured by inner product), are given higher feature values in the final solution, placing them farther out in **WAS** space as measured by Euclidian distance. This fact will have important implications for the way in which the **WAS** vectors are incorporated in a Bayesian analysis, and the way in which word frequency is treated, as described below.

Orthographic features

For convenience, the physical features of words were represented only and simply in terms of orthographic features. The role of physical aspects such orthography is emphasized in this research because the orthographic similarity of test words to studied words was varied in one of the experiments in this paper. In principle, the present modeling effort could easily be extended to include other aspects of words such as phonology, or font, style, size and capitalization.

In this research, of the many possible ways to encode orthography, a simple representational scheme was chosen that is based on the probabilities of letters occurring in words. First, the distribution of letter frequencies was computed by counting the occurrences of letters in a large lexicon of CELEX (Burnage, 1998). Let us denote the jth most frequency letter in the alphabet with Q_j and the relative frequency of Q_j with h(Q_j). For example, the most frequent letter in our frequency count is “e” so Q_1=“e” and we calculated h(Q_1) = .0997. The idea is to code words with the ranks of the letter frequencies as illustrated in Figure 3. With this representation, the word “bear” would be encoded with the four features 16-1-2-3 and the word “rex” with the three features 3-1-25.

LETTER	FREQ.	CODE
e	0.0997	1
a	0.0823	2
r	0.0795	3
…	…	
b	0.0247	16
p	0.0235	17
k	0.0197	18
…	…	
x	0.0025	25
q	0.0017	26

Figure 3. Illustration of the representation for orthography. Letters are encoded with the rank of the frequency with which the letter appears in a large lexicon.

The base rates of feature values h(Q_j) are assumed to be known to the system. Based on these base rates for the features, the memory model can predict word frequency effects. High frequency words consist on average more of high frequency features while low frequency words consist on average more of low frequency features. A match of a low frequency feature between a test word and a memory trace provides highly diagnostic evidence in favor of a match, whereas a match of high frequency features is more likely to have occurred by chance and therefore provides less evidence. These differences in diagnosticity present one way in which the model can predict word frequency effects (similar arguments apply in principle to diagnosticity of semantic features and word frequency, but the peculiarities of **WAS** do not lend themselves to the appropriate Bayesian analysis--see below).

<u>Episodic storage</u>

Study of words leads to episodic traces in memory, separately for each word. The traces in memory are error prone and potentially incomplete copies of the semantic and orthographic feature vectors. With probability u, a semantic/orthographic feature is stored in a trace. If a feature is not stored, it is marked as missing and cannot be part of the retrieval process. A high probability u leads to relatively complete traces in memory whereas a low probability u leads to weak traces in memory.

In the original REM model, the feature values representing words were discrete. In this model, the orthographic feature values are discrete and the semantic feature values are continuous, so different processes are used to add noise in the storage process. For the discrete orthographic features, the parameter c determines the probability that feature values are copied correctly into the episodic trace. If a feature is not copied correctly, it is sampled from the distribution of feature values. Therefore, if it is not copied correctly, the most likely value to be stored is “1”, next most likely value is “2”, and so

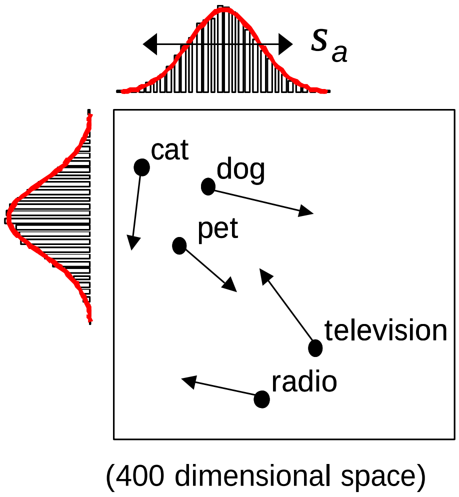


Figure 4. Illustration of the storage process for semantic features. Normally distributed noise with standard deviation σ_n was added to each feature (or dimension).

forth.

For the continuous semantic features, normally distributed noise is added for each feature value as illustrated in Figure 4. The parameter σ_n, the standard deviation of the noise distribution determines the amount of noise in the storage process for semantic features. In all, three parameters, u, c and σ_n determine the storage process. In light of the peculiar properties of **WAS**, one might wonder whether it is sensible to add constant noise to all feature values. In principle this is an excellent question. In practice, the relative placement of high and low frequency items in **WAS** caused us to normalize all semantic vectors by their length (see below), thereby placing all words on a hypersphere, and thereby making the constant noise assumption plausible.

Calculating Familiarity

The recognition decisions are based on Bayesian principles where the log odds is calculated that the probed word is old over new:

$$j = \log \frac{P(probe\ is\ old)}{P(probe\ is\ new)} \qquad (1)$$

In REM, binary recognition decisions "old" and "new" are made when the log odds is bigger than zero and smaller than zero respectively. In this research, we will model not binary recognition decisions, but recognition judgments that lie on a scale. For this purpose, we first took the log of the odds, thereby making the log odds distributions at least roughly normal for both targets and distractors (see Shiffrin & Steyvers, 1997). These log odds could then be transformed into a judgment scale.

In the model, if the probe is a target, one of the traces is a result of storing that probe, but which trace is not known to the system. If the probe is a distractor word, none of the traces are the result of storing that probe. Because the storage process is made noisy, it can only be determined probabilistically whether one of the traces match the probe. In the appendix of Shiffrin and Steyvers (1997), it was shown with Bayesian principles how to calculate the odds that the probe is old over new. The calculations use the available information: the matching of the features of the probe to those of the stored features in each memory trace. First, the odds is expressed as a sum of the likelihood ratio's, λ_i of the individual trace i matching the probe, divided by the number of traces, n:

$$j = \frac{1}{n}\sum_{i=1..n} l_i \qquad (2)$$

The likelihood ratio λ_i, expresses the ratio of the probability that the test probe was stored in trace i over the probability that the test probe was not stored in trace i.

To combine evidence from orthographic feature matches, and semantic feature matches, one simply multiplies likelihood ratios:

$$l_i = l_i^s\, l_i^o \qquad (3)$$

where λ^s_i and λ^o_i denote the likelihood ratios calculated for the semantic and orthographic contents in memory respectively.

As with the discrete features of the original REM model, the number of matching and mismatching features between the probe and trace are used to calculate the likelihood ratio's for orthographic features:

$$l_i^o = \prod_{k\in N_i}(1-c)\prod_{k\in M_i}\frac{c+(1-c)h\left(V_{k,i}^o\right)}{h\left(V_{k,i}^o\right)} \qquad (4)$$

The sets N_i and M_i index the set of features of trace i that match and mismatch the probe respectively. The variable $V^o_{i,k}$ refers to the kth orthographic feature stored in the ith trace in memory. The parameter c and function h(V) were introduced earlier. The parameter c determines the probability that features are stored correctly. The function h(V) is the distribution of orthographic feature values that was determined by the relative letter frequencies of letters appearing in words in a large lexicon.

The likelihood ratio's are calculated for every trace in memory. Therefore, the number of matching and mismatching orthographic features is calculated for every probe-trace comparison. Because words differ in length, it becomes an issue of how to align probe and trace features in case there is a length mismatch. There are various solutions to this problem. Here, the best alignment was chosen for each probe-trace comparison; 'best' is defined in terms of the least number of mismatches.

For a continuous metric space in which similarity is inversely related to distance, it would be sensible to use the absolute difference between two features values as a way to measure the degree of match between features. However, in **WAS** high frequency words, which are highly similar, and have common features, are placed in the outskirts of the space (i.e. they have larger feature values). For such a representation, we could find no way to instantiate or approximate a sensible Bayesian implementation. We therefore normalized all vectors in **WAS** by dividing all feature values for a word by that word's vector length[1]. This placed all words on the surface of a hypersphere, and similarity is inversely related to distance on this hypersphere. For this new representation, it is plausible to measure degree of match by absolute difference between feature values (although, as discussed below, an unfortunate consequence of this change is the elimination of feature frequency differences between words of different frequency).

Based on Bayesian principles, it can be shown that the likelihood calculation for the semantic features defined in this way is:

$$l_i^s = \prod_{k=1..K} \frac{f\left(\left|V_{k,i}^s - W_k^s\right|\right)}{g(V_{k,i}^s)} \quad (5)$$

The variable $V^s_{i,k}$ refers to the kth semantic feature stored in the ith trace in memory, W^s_k refers to the kth semantic feature in the probe and K refers to the number of semantic features (K=400). The function f is the probability mass distribution of the normal distribution with standard deviation σ_n. The numerator is the probability density of the observation assuming the probe word had been stored in trace i, and the denominator is the density under the assumption that trace i encodes some other word[2]. The ratio gives the ratio of evidence for feature k, and the product of these gives the likelihood ratio for the ith trace.

Recognition and Similarity Judgments

It is assumed that both semantic and orthographic features are used when making recognition judgments, whereas only semantic features are used when making similarity judgments. The system in Equations (2)-(5) determines how the familiarity values for recognition judgments are calculated. In order to calculate the familiarity values for the similarity judgments, orthographic features were deleted, by changing Equation 3 to:

$$l_i = l_i^s$$

In order to distinguish the log odds calculated for the recognition and similarity judgments, they will be referred to as $\varphi_{\text{recognition}}$ and $\varphi_{\text{similarity}}$ respectively.

Word frequency effects

Word frequency effects might well be due to feature frequency differences, at least in part. The present model incorporates this factor only for orthographic features, and hence only for recognition judgments, not similarity judgments. To construct a sensible Bayesian analysis for **WAS**, it was necessary to normalize the vector lengths, placing all words on a hypersphere, and eliminating feature frequency differences between high and low frequency words. This greatly diminished word frequency effects for recognition (they are based only on orthographic diagnosticity) and eliminated them for similarity judgments.

It should be emphasized that these normalization changes we have made to **WAS** are technical in nature, and it remains quite possible that word frequency effects are due in substantial part to feature frequency diagnosticity. If, for example, it had been possible to use multidimensional scaling for a database as large as that in the Nelson et al. (1998) norms, it is quite possible that the resultant space would cluster high frequency words closer than low frequency words, and would place the features of high frequency words closer than those of low frequency words to the mean values on each dimension. Due to the computational demands of applying a multidimensional scaling procedure on the norms, it was not presently possible to carry out such analyses, unfortunately.

Be this as it may, real data requires the prediction of word frequency effects. Because a feature frequency basis for such predictions is not available (except for the orthographic component of recognition judgments), we decided to base such predictions on another factor, the enhanced recency and greater number of contexts for high frequency items: does the test word appear familiar because it was studied, because it was seen recently or because the current context matches one of the many possible contexts in which the high frequency word appears? Dennis and Humphreys (submitted, 1998) constructed a Bayesian model that explained word frequency effects based on this factor. However, adding such a system to our present modeling effort would add a great deal of complexity and take us quite far afield. We decided instead to approximate the results of such a system in the following descriptive way, a way that would incorporate word frequency effects, and also produce mirror effects. A reference value, γ, was assumed toward which all calculated (log) odds are regressed (i.e. squeezed). The amount of regression is higher for high frequency words, according to the following equations (the values of α are between 0 and 1):

$$j'_{recogniti\mathbf{a},F} = a_F j_{recogniti\mathbf{a},F} + (1-a_F)g \quad (6)$$
$$j'_{similarity,F} = a_F j_{similarity,F} + (1-a_F)g$$

The value of α_f was made a monotonically decreasing function of the word frequency F of the probe:

$$a_F = \left(1 - 1/\sqrt{b}\right) + 1/\sqrt{b+F} \quad (7)$$

A zero word frequency is mapped to α=1. Higher word frequencies lead to lower α values where the falloff is determined by scaling parameter b. The parameter γ in Equation (6) determines the centering of the mirror effect for word frequency. Suppose the mean distractor and target familiarity is lower and higher than γ respectively. Compared to low frequency distractors, the familiarity will be increased toward γ for high frequency distractors. Compared to low frequency targets, the familiarity will be decreased toward γ for high frequency targets. Increasing the value of γ, leads to an increasing frequency effect on distractors but decreasing effect on targets. Decreasing the value of γ, leads to a decreasing frequency effect on distractors, but increasing effect on targets. Thus equations 6 and 7 represent a purely ad hoc, but fairly simple, method by which to approximate the effect of a recency/context factor for word frequency.

Predicting Individual Word Differences.

The model utilizes the particular words for a given trial, and makes predictions for particular test items, based on the orthographic and similarity relations among the various words. The ability of the model to capture the variability in performance due to individual word differences was measured by the correlation between observed and predicted judgments for individual words. The correlational analyses were performed in two ways: single and multiple conditions.

In the single condition analyses, only words from a single condition were included for each correlational analysis: Significant correlations indicate the model explains significant parts of variance due to individual word differences. This procedure is somewhat limited because some conditions do not contain enough words to draw strong statistical conclusions. In the multiple condition analyses, words from different sets of conditions were pooled to calculate the correlation. However, any resulting correlations are due to a mixture of within and between condition effects, so no conclusions can be drawn concerning the gains due to individual word predictions. The situation is illustrated in Figure 5: the horizontal axis shows some measure of similarity between test word and studied words. Only in Figure 5a is there a within condition correlation that could be interpreted as indicating additional predictability due to consideration of similarities between particular

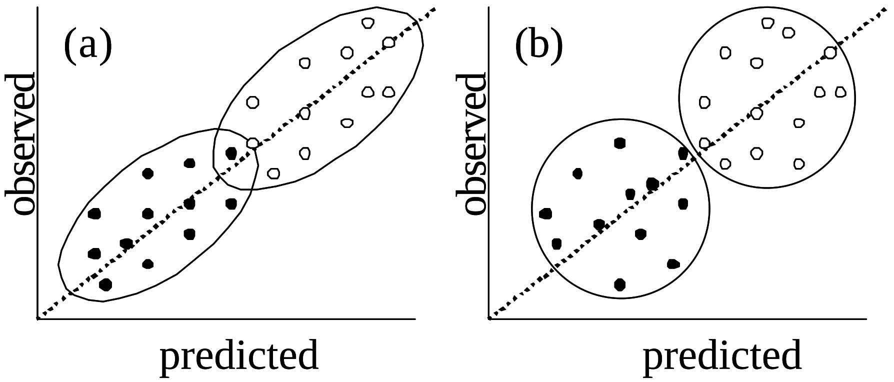


Figure 5. Two possibilities situations for calculation correlations between observed and predicted results for individual words when words from two different conditions are pooled. The dots represent different words and the color differences represent condition differences. In (a), part of the correlation between observed and predicted results is due to the capturing part of the within as well as the between condition variability. In (b), the correlation is solely due to between condition differences.

words. Both panels show substantial between condition correlations.

Overview of Experiments

We present three experiments in which distractor similarity, the length of studied categories and the directionality of association between study and test words were varied. The comparison of the results for the recognition and similarity judgments is important to investigate the interplay between semantic and physical features in recognition memory. The experiments address five basic predictions of the memory model:

(1) Testing distractor words that are increasingly semantically similar to studied words will lead to increasingly higher false alarm rates. This is simply a result of the model being a global familiarity model: it computes the overall match between the probe and contents of memory. Since semantic similarity is determined by the semantic space of WAS, for a given set of study words, the model can make specific predictions about which words will lead to what level of false alarms relative to other words. This prediction was addressed in Experiment 1, 2, and 3.

(2) Increasing the orthographic similarity between a distractor word and the stored orthographic contents in memory will increase the false alarm rates. This prediction was addressed in Experiment 2.

(3) The difference between recognition and similarity judgments was assumed to be due to a reliance on different sources of information. For similarity judgments, only semantic features were used while for recognition judgments, both semantic and physical features such as orthographic features were used. Therefore, the effect of semantic similarity of distractors should have a larger effect on similarity judgments than recognition judgments. Also, there should be no effect of orthographic distractor similarity on similarity judgments (the similarity judgments imply semantic similarity). These predictions were addressed in Experiment 2.

(4) The model should capture part of the variability in performance due to individual word differences, above and beyond the variability due to between condition differences. This prediction was addressed in all three experiments.

(5) A word frequency effect is predicted: low frequency words have higher hit rates and lower false alarm rates than high frequency words. This prediction is addressed in all three experiments.

Experiment 1

This experiment tests the ability of the model to predict the false alarm rates to semantically similar distractors. The closer in **WAS** are distractors to studied words, the more false alarms should be produced. Four groups of distractors were created (labeled A, B, C, and D) that were monotonically decreasing (from A to D) in their semantic similarity to studied words. Each group has subgroups of low and high frequency words. Word frequency was varied in this experiment to investigate the interaction between distractor similarity and distractor word frequency.

Method

Design and Subjects. For the distractors, the design formed a 4 x 2 factorial, with word frequency (low, high) and distractor similarity (four groups A, B, C, and D that were increasingly less similar to studied words) manipulated within subjects. For targets, only word frequency (low, high) was manipulated as a within-subject factor. Thirty-five students from Indiana University who were enrolled in introductory psychology courses participated in exchange for course credit.

Materials. Appendix A shows the words from this experiment for each level of word frequency and distractor similarity. All words were selected from the Nelson et al. (1998) free association norms. Word frequency was operationally defined by the number of times the word was produced as an associate in the norms of Nelson et al. (1998). We defined low frequency words as words that were produced by less than 10 of the 5018 total cues of the norms. High frequency words were defined as words produced by 10 or more cues. The low and high frequency words in the experiment were produced by an average of 4.2 (SD=3.4) and 30.3 (SD=17) cues respectively. We also measured differences of the resultant groups in the Kucera and Francis frequency count, which is the traditional way to measure and define word frequency. The low and high frequency words had median Kucera and Francis frequency counts of 5 (SD=9.2) and 28 (SD=126) respectively. Therefore, the low and high frequency words had both different production counts and Kucera and Francis frequency counts.

On the basis of 18 randomly selected prototype words, 18 categories were created. Within **WAS**, the four most similar low frequency words and the four most similar high frequency words to each of the prototype words were selected. Similarity between two words was computed by the inner product of the two vectors in **WAS (**In this method section, when we refer to WAS, we refer to the vectors whose lengths were not normalized). The 4 low and 4 high frequency words of each of 18 categories served as study words in the experiment.

The distractor words varied in both word frequency and similarity to the 18 study categories.

For each frequency level, four similarity groups were created that varied in the similarity to studied categories, from very high (group A) to very low (group D). We manipulated distractor similarity by varying the degree of similarity of words to specific categories on the study list rather than to all the words on the study list. Distractor similarity was operationally defined by using the mean WAS similarity of a distractor word to the words from a specific study category. For each study category, the mean similarity of each of the 5018 words from the norms to the category words was computed (excluding all study words). Four high frequency groups, and four low frequency groups of similarity were created by selecting words with similarity measures ranging between .10 - .45, .05 - .10, .02 - .05, and .0018 - .0045 respectively. Averaged over word frequency, the average similarity of the four groups was respectively .1853, .0869, .0354, and .0027. In other words, the words from groups A to D decreased monotonically in their mean similarity to categories on the study list.

Procedure. An experimental session consisted of one study-test cycle. Participants were instructed prior to the presentation of the study words to remember the words on the study list. Each word was displayed in the center of the computer screen for 1.3 s. of study. The category words were presented one after the other until all the words from a category were presented and the next category was selected. The order of words within a category as well as the order of categories on the study list was randomized for each participant. The study list consisted of 144 study trials, including the 18 categories of 8 items each.

The procedure of Brainerd and Reyna (1998) was changed in two ways. In their studies, the two memory judgments were varied between groups. In our experiments, each test item required two memory judgments. Second, instead of binary “yes”, “no” judgments, our participants were asked to give judgments on a six point scale. After study, participants read detailed instructions. Participants were informed that they would give two ratings for each test word: a recognition rating and a similarity rating. For the recognition rating, participants were instructed to rate how confident they were that a test word had been studied by utilizing a 6-point scale (a 1 indicated high confidence that the word had not been studied and a 6 indicated high confidence that the word had been studied). They were also instructed to give low ratings to distractor words that were similar to the studied categories, if that test word was not an exact match to a studied word. For the similarity rating, participants were instructed to rate how confident they were that words similar in meaning had been studied by utilizing a 6 point confidence scale (a 1 indicated high confidence that no similar words had been studied and a 6 indicated high confidence that words similar in meaning had been studied). They were also instructed to give high similarity ratings if the test word had in fact been studied.

There were a total of 100 test items. Of the test items, 28 were targets, and 72 were distractors. Of the 28 target items, 14 were low frequency and 14 were high frequency words. The target items were chosen randomly from the pool of study words with the constraint that each category was tested at least once and at most twice. The 72 distractor items consisted of equal numbers of items from the 4 distractor groups A, B, C, and D. Each distractor group consisted of an equal number of low and high frequency distractors. The distractor items were chosen randomly (sampling equally from low and high frequency groups) from the pool of distractor words with the constraint that each category was tested exactly four times.

Results

For each participant, the confidence ratings for the recognition judgments were converted to z-scores by subtracting the mean and dividing by the standard deviation of all the recognition confidence ratings for that participant. The z-scores were then averaged over participants to get the overall z-scored ratings for a given condition. The same procedure was applied to the confidence ratings of the similarity judgments. The conversion to z-scores has the advantage of normalizing for idiosyncratic uses of the 6 point confidence scales. For example, some participants use one end of the scale more than the other and some participants give wider ranges of ratings than others. By subtracting the mean and dividing by the standard deviation of the ratings, much of the participant specific variance was eliminated. Note that positive recognition and similarity z-scores indicate more than average confidence that the item is old and similar, respectively. Similarly, negative recognition and similarity scores indicate more than average confidence that the item is new and dissimilar respectively.

We also computed d’ as a measure of sensitivity: the degree to which targets and distractors were discriminated. In order to compute d’, we first computed for each participant the median confidence ratings for the recognition judgments and similarity judgments separately. The median confidence rating was used a criterion below which the response would be scored as a “no” judgment and above which the response would be scored as an “yes” judgment. The probability of responding “yes” for targets and distractors then served as hit and false alarm rates for a given condition in order to compute d’ for each participant separately. Repeated measures analyses of variance (ANOVA’s) were conducted on the z transformed recognition and similarity judgments as well as the sensitivity measures. In each analysis, the Type I error rate was set at .05.

Recognition judgments. The means and standard errors of the recognition and similarity z-scores for the high and low frequency targets and for the low and high frequency distractors in the four similarity groups are shown in Figure 6. This figure shows that participants rated the distractor items from groups A to D as increasingly less “old”. This effect is observed for both low and high frequency items. The figure also shows that low frequency distractors are rated more as “new” than high frequency distractors whereas low frequency distractors are rated as slightly more “old” than high frequency distractors. For distractors, the effect of similarity was significant [F(1,34)=103, MSE=.0618] as well as the effect of word frequency [F(1,34)=47.1, MSE=.0872]. The interaction of both effects was not significant [F(1,34)=1.71, MSE=.0776, p<.20]. For targets, the effect of word frequency was not significant [F(1,34)<1].

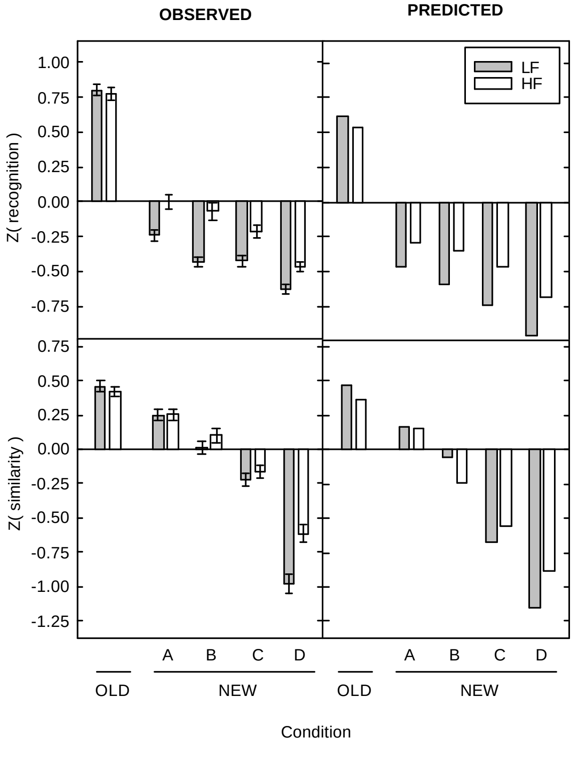


Figure 6. Observed and predicted results of Experiment 1.

Table 1 lists the mean d’ results as well as the standard error of d’ based on several target and distractor condition comparisons. The results show that participants are increasingly more able to discriminate between old items and new items from groups A to D. Also, sensitivity for low frequency items is higher than for high frequency items. The effect of similarity on sensitivity was significant [F(1,34)=48.5, MSE=.409] as well as the effect of word frequency [F(1,34)=13.0, MSE=.915] while the interaction was not significant [F(1,34)<1].

Similarity judgments. The similarity ratings decreased progressively from group A to group D distractors. The effect of distractor similarity was significant [F(1,34)=207, MSE=.194]. Although the effect of word frequency on distractors was significant [F(1,34)=11.48, MSE=.101], Figure 6 shows that the effect is caused mainly by the differences between low and high frequency items of group D. Paired sampled t-tests confirm that only this group showed a significant word frequency effect [t(34)=4.2]. Removing this group from analysis led to non-significant effects of word frequency [F(1,34)=1.72, MSE=.0778, p<.2]. For targets, the effect of word frequency was not significant [F(1,34)<1].

The sensitivity results for the similarity ratings follow the same pattern as the recognition ratings: the ability to discriminate between old and new items increases with decreasing distractor similarity. This effect was significant [F(1,34)=170, MSE=.568]. The effect of word frequency was marginally significant [F(1,34)=3.87, MSE=.827, p<.057] and became non significant after removing group D distractors [F(1,34)=1.27].

Number of ratings per word. Each of the 35 participants was tested on different subsets of words available for study and test. Each of the target words from the pool of 144 words was rated by a median of 7 participants (SD=2.3). Each of the distractor words from the pool of 144 words was rated by a median of 18 participants (SD=2.8). Because the target words were judged by only few participants, they were excluded from the correlational analyses of observed and predicted results that will be discussed shortly.

Discussion.

The results show three clear patterns. First, the distractors that are increasingly less similar to studied categories, where similarity is defined by inner products in WAS, are rated as more “new” and “dissimilar”. This suggests that the semantic space can be helpful in predicting the false alarm rates of

Table 1
Sensitivity results (d') for Experiments 1,2 and 3

	Observed				Predicted	
	Recognition		Similarity		Recognition	Similarity
Comparison	M	StdErr	M	StdErr	M	M
			Experiment 1			
low frequency						
OLD vs. NEW-A	1.2	0.1	0.38	0.1	1.28	0.32
OLD vs. NEW-B	1.52	0.13	0.73	0.1	1.47	0.6
OLD vs. NEW-C	1.52	0.11	1.09	0.15	1.82	1.38
OLD vs. NEW-D	1.99	0.14	2.14	0.15	2.31	2.22
high frequency						
OLD vs. NEW-A	0.85	0.12	0.23	0.09	1.18	0.17
OLD vs. NEW-B	0.99	0.14	0.54	0.11	1.27	0.71
OLD vs. NEW-C	1.14	0.11	1.03	0.12	1.55	1.29
OLD vs. NEW-D	1.6	0.14	1.69	0.18	2.12	2.09
			Experiment 2			
blocked - semantic						
OLD-3-PRO vs. NEW-3-PRO	1.07	0.24	0.22	0.27	1.68	0.5
OLD-7-PRO vs. NEW-7-PRO	1.58	0.24	0.15	0.24	2.32	0.52
OLD-3-EXE vs. NEW-3-EXE	1.57	0.16	0.74	0.16	1.93	0.8
OLD-7-EXE vs. NEW-7-EXE	1.22	0.17	0.59	0.14	1.86	0.47
blocked - orthographic						
OLD-3-PRO vs. NEW-3-PRO	0.55	0.18	0.46	0.19	1.6	1.15
OLD-7-PRO vs. NEW-7-PRO	1.21	0.2	0.67	0.23	1.62	1.09
OLD-3-EXE vs. NEW-3-EXE	1.06	0.13	0.73	0.15	1.57	1.38
OLD-7-EXE vs. NEW-7-EXE	0.85	0.1	0.6	0.1	1.35	1.39
spaced - semantic						
OLD-3-PRO vs. NEW-3-PRO	1.09	0.26	0.39	0.27	1.33	0.52
OLD-7-PRO vs. NEW-7-PRO	0.75	0.3	0.07	0.23	1.28	0.48
OLD-3-EXE vs. NEW-3-EXE	1.25	0.13	0.67	0.14	1.44	0.6
OLD-7-EXE vs. NEW-7-EXE	1.05	0.19	0.51	0.14	1.46	0.66
spaced - orthographic						
OLD-3-PRO vs. NEW-3-PRO	0.61	0.23	0.76	0.21	1.35	1.01
OLD-7-PRO vs. NEW-7-PRO	0.49	0.24	0.64	0.28	1.16	0.81
OLD-3-EXE vs. NEW-3-EXE	0.71	0.12	0.7	0.11	1.31	1.17
OLD-7-EXE vs. NEW-7-EXE	0.69	0.11	0.54	0.13	1.27	1.27
			Experiment 3			
OLD-A vs. NEW-LF	1.88	0.1	1.14	0.12	1.78	0.91
OLD-B vs. NEW-HF	1.24	0.11	0.72	0.12	1.2	0.76
OLD-A vs. NEW-F	1.11	0.13	0.11	0.12	0.52	0.03
OLD-B vs. NEW-G	2.02	0.14	0.21	0.15	1.62	0.34
OLD-C vs. NEW-H	0.82	0.09	0.05	0.09	1.02	0.29

distractor words. Second, word frequency had the predicted effect on recognition judgments for distractors: high frequency distractors were rated as more “old” than low frequency distractors. Interestingly, the effect of frequency on similarity judgments was less pronounced. Apart from group D distractors, there was only a small increase in the “old” ratings for high frequency distractors compared to low frequency distractors. Third, the participants can distinguish between recognition and similarity ratings. When the results for similarity and recognition judgments are compared, the difference between group A distractors and targets is much smaller for the similarity ratings than for the recognition ratings. This indicates that participants are following instructions because they were instructed to give high similarity ratings to test words that were similar to studied words regardless of whether the test words were studied or not.

Model Fits of Experiment 1

The model as outlined in the Introduction was applied to Experiment 1. The same study and test words were used in the model as in the experiment. In total, there were four parameters to model the experiment. The two storage parameters, c (0.2) and σ_n (0.25) determined the amount of storage noise for orthographic and semantic features respectively. The parameter γ (3.0) determined the centering for the word frequency effect and the b parameter (5.0) was used as a parameter to scale the word frequency effect. These were all the parameters that were needed to generate predictions. No iterative techniques were used to find the “best” parameter settings to optimize the fit between observed and predicted results. Only a handful of parameter setting were tried until the predicted results showed (most of) the desired qualitative pattern of results[3].

Recognition and Similarity Judgments. Figure 6 shows the predictions of the model obtained by simulating 100 participants. In the experiments, the recognition and similarity judgments were Z-transformed. In the modeling, the $\varphi'_{similarity}$ and $\varphi'_{recognition}$ familiarity values were also Z-transformed. The model results capture three basic trends in the data. First, a monotonic decrease in the “old” ratings was predicted for conditions A to D. On the one hand, this is not surprising because conditions A to D contained words that are semantically increasingly dissimilar according to the semantic space formed by WAS. However, this does suggest that the word vectors in the semantic space are organized appropriately and gives the semantic space some psychological plausibility. Second, the difference between recognition and similarity judgments is correctly predicted. The difference between targets and the semantically closest distractors (group A) is predicted to be much smaller for the similarity than recognition judgments. Recognition judgments use orthographic features to help distinguish targets from semantically similar distractors. Third, word frequency effects were predicted mainly because of the descriptive component in the model that squeezed familiarity values towards the center of scale to a degree dependent on word frequency. This approximation was employed to mimic the effects of recency and context noise; although feature frequency effects ought to have operated as well, the normalizing of **WAS** eliminated the possibility of including this component in the model.

Sensitivity. The d’ results for the model’s predictions were generated in the same way as in the experiments. For each simulated participant, a criterion for the recognition and similarity judgments was determined by taking the median of the $\varphi'_{recognition}$ and $\varphi'_{similarity}$ familiarities respectively (over all conditions). These criteria specify the midpoint of the recognition and similarity scale above and below which lie 50% of the judgments. The sensitivities were then calculated on the probabilities of responding above the criterion for targets and distractors respectively.

The predicted d’ results (Table 1) show the same pattern as the observed d’ results. The sensitivity for low frequency words is higher than for high frequency words. This is a direct consequence of the familiarity values for high frequency target and distractor words being squeezed toward the center of the familiarity scale. The sensitivity monotonically increased from group A to group D because of the monotonically decreasing false alarm rates for these groups.

Individual Word Correlations. Table 2 shows correlations for the predicted and observed Z-scores of individual words with words from single as well as multiple conditions. The first column shows which conditions were used in the calculating the correlation. The second column shows the number of words in the comparison. The next three columns show the results from the correlational analyses for the recognition ratings while the last three columns those for the similarity ratings. In the column “original”, the correlation value is shown with potential markers for statistical significance. The “scrambled” column shows the correlation value under a procedure in which the order of words within each condition is scrambled so that the resulting correlational value can only be attributed to predicted between condition differences and not to predicted individual word differences within condition[4]. For correlations that only involve words from a single condition, the scrambled correlational

Table 2

Correlations between predicted and observed z-scores for recognition and similarity ratings

			Recognition			Similarity		
Groups	N	MC[a]	original	scrambled	diff[b].	original	scrambled	diff[b].
Experiment 1[c]								
all distractors	144	y	.50***	.35***	**	.67***	.56***	**
all LF distractors	72	y	.37***	.26**		.70***	.61***	*
all HF distractors	72	y	.52***	.29***	**	.64***	.48***	**
A-LF	18	n	0.07	0		0.18	0	
A-HF	18	n	0.06	0		.39*	0	*
B-LF	18	n	0.22	0		.38*	0	*
B-HF	18	n	.43**	0	**	.32*	0	*
C-LF	18	n	.36*	0	*	.36*	0	*
C-HF	18	n	.56***	0	***	.40*	0	*
D-LF	18	n	0.2	0		0.13	0	
D-HF	18	n	0.15	0		0.12	0	
Experiment 2								
all	1234	y	.63***	.58***	***	.49***	.44***	***
all targets	562	y	.12***	0.05	*	.21***	.19***	
all distractors	672	y	.37***	.18***		.44***	.31***	***
target exemplars	371	y	.11**	0.01	**	.22***	.19***	
target prototypes	191	y	.13**	.13**		.19***	.22***	
related exemplar distr.	384	y	.27***	0.05	***	.45***	.28***	***
related prototype distr.	192	y	.26***	.11*	**	.32***	.30***	
Experiment 3								
all	180	y	.86***	.85***		.63***	.64***	
all targets	60	y	0.05	0.04		0.16	0.03	
all distractors	120	y	.62***	.56***		.64***	.63***	
all related distractors	60	y	.63***	.62***		0.01	-0.01	
all unrelated distractors	60	y	.63***	.51***	*	.43***	.38***	
A	20	n	0.09	0		0.19	0	
B	20	n	0.17	0		0	0	
C	20	n	.44**	0	**	.44**	0	**
F	20	n	0.24	0		0.12	0	
G	20	n	0.23	0		0.02	0	
H	20	n	0.08	0		0.06	0	
LF	30	n	.47***	0	***	0.03	0	
HF	30	n	0.05	0		.33**	0	**

Notes

*** p<.01 ** p < .05 * p < .10

a. "y" for multiple conditions involve correlations for words of multiple conditions

b. This columns indicates whether the difference in correlation for original and unscrambled words is significant

c. The correlations for words in the target conditions are not shown because there were not enough participants that rated each individual target word.

value is by definition zero because no between condition differences can be defined. The "diff" column lists the statistical significance of the difference between the original and scrambled correlational values. If such a difference is found to be significant, it means that a significant part of the variability in the observed results within conditions can be explained in the model on the basis of individual word differences. In the present experiment, of course, the conditions themselves involve variations of similarity along the same dimensions as those operating by chance within condition. Thus the two correlational analyses are in a sense redundant and ought to give rise to the same conclusions.

Table 2 shows that the correlations are higher for the similarity ratings than for the recognition ratings. This is interesting because for similarity judgments, the variability in the model is only due to semantic features while for the recognition judgments, an additional source of variability is provided by the orthographic features.

For five out of eight single condition groups, the correlation was higher than .3. This is a very small correlation but it should be kept in mind that in these analyses, the range of distractor similarities within condition was limited: because the stimuli were chosen approximately to equate similarities within condition, the differences in similarities that remained were accidental and limited in scope. Also, in each of these conditions, only 18 words were part of the correlation, so that statistical significance was harder to reach than for the multiple condition correlations. More impressive are the correlations for words from multiple conditions. When all low frequency distractors or all high frequency distractors were part of the correlational analysis, the correlation for the similarity ratings was moderately high (>.6) and higher than in the scrambled procedure. This indicates that the memory model with the derived semantic similarity relationships in WAS can predict part of the variability in similarity judgments due to individual word differences, both across and even within condition.

Parameters. The four parameters[5] used to generate predictions for this experiment were set at: σ_n =.25, c=.2, b=5, γ=3. Note that the noise distribution for semantic features has a standard deviation five times larger than the standard deviation of all semantic feature values in WAS (.0484). Such a large noise value is needed because there are 400 diagnostic feature values which together provide a good deal of information even in the face of a great deal of feature noise.

It might be expected that appropriate values for γ should be around 0 because a log odds of 0 should be the center of the familiarity scale for Bayesian models (see Shiffrin & Steyvers, 1997). However, we violate a key assumption of the simple Bayesian derivation: the study words were not sampled randomly from the pool of all possible study words. Instead, we sampled groups of semantically similar words. Therefore, the log odds distributions for both targets and distractors were not centered around zero, requiring that the centering for the mirror effect be placed on familiarity values higher than zero. The particular value chosen also allowed the model to handle the fact that word frequency affected distractors more than targets.

Experiment 2

Several studies have shown that hits and false alarms go up monotonically with the number of same-category items on the study list (Hall & Kozloff, 1970; Hintzman, 1988, Robinson & Roediger, 1997; Shiffrin et al., 1995). For example, if the study list contains fruit words (e.g. apple, pear, banana, etc.), the hit rate for a studied fruit word and the false alarm rate to new fruits will typically increase with the number of fruit words studied. Hintzman (1988) and Shiffrin et al. (1995) have given quantitative accounts of this category length effect solely on the basis of global familiarity: a test word that is related to more traces in memory results in higher global familiarity.

Shiffrin et al. (1995) have argued that in their study, it is unlikely that related unstudied category words were thought of during study or were activated by a spreading activation mechanism, because all category words were randomly spaced over the study list. It is more likely that the IAR mechanism or a spreading activation account plays a role when the category words are studied in a blocked fashion. It is hard to imagine that participants will not think about the prototype “fruit” when fifteen fruit words are studied one after the other. Several studies have investigated the effect of studying the category words in a blocked or spaced fashion (Agostino, 1969; Mather, Henkel, & Johnson, 1997; Toglia, Hinman, Dayton, & Catalano, 1997). Mather et al. reported that both the hit rate for studied words and false alarm rates for unstudied prototypes were higher in the blocked presentation condition but that false alarm rates for unrelated distractors were lower in the blocked study conditions.

While both the category length effect and the blocked/random effect have been investigated, the interaction of these effects have not been explored yet. The goal of this experiment is to investigate the effect of study presentation (blocked/random) and category kind (semantic or orthographic/phonological) on the category length effect.

Method

Design and participants. The design formed a (2 x 2 x 2 x 2) + 2 mixed factorial design. Study presentation (blocked vs. spaced) was varied between subjects. Category length (3 or 7), category type (orthographic or semantic) and category membership (prototype or exemplar) was varied within subjects. Two distractor conditions were added, containing words that were unrelated to studied categories

Table 3
Within subject conditions of Experiment 2

Condition	Target	Category Kind	Category Length	Prototype or Exemplar	#tested
1	Y	S	3	P	4
2	Y	O	3	P	4
3	Y	S	3	E	8
4	Y	O	3	E	8
5	Y	S	7	P	4
6	Y	O	7	P	4
7	Y	S	7	E	8
8	Y	O	7	E	8
9	N	S	3	P	4
10	N	O	3	P	4
11	N	S	3	E	8
12	N	O	3	E	8
13	N	S	7	P	4
14	N	O	7	P	4
15	N	S	7	E	8
16	N	O	7	E	8
17[b]	N	S & O	0	E	8
18[c]	N	S & O	0	P	8

Note: Y=yes, N=no; S=semantic category, O=orthographic category; P=prototype, E=exemplar

a. When a prototype is tested as a target it was on the study list

b. These distractor words were drawn from the pool of exemplar words of unstudied categories

c. These distractor words were drawn from the pool of prototype words of unstudied categories

(essentially 0 category length) and that were either drawn from the pool of unused prototype or exemplar words. Table 3 summarizes the within subject conditions in this experiment. Thirty-seven participants were assigned to the blocked condition and thirty-four participants to the spaced condition. The participants were drawn from the same pool of participants of Experiment 1.

Materials. The words from this experiment are listed in Appendix B. All words were part of the Nelson et al. (1998) norms. Twenty four words were pseudo- randomly selected from the pool of words to serve as prototypes for the semantic categories (these were chosen by hand so that they seemed to be plausible candidates for category prototypes). For each of the 24 prototype words, 9 exemplar words were chosen that were most similar to the prototype words in the WAS space. The exemplars were picked with the constraint that the words were not used for other categories and that the words from the same word form were not used (e.g. choosing “egg” and “eggs” as exemplar words for the same category was not allowed). 24 orthographic categories were created by pseudo-randomly selecting 24 prototype words from the pool of words. For each of the 24 prototype words, 9 exemplar words were selected that differed in one or two letters from the prototype word.

Procedure. Participants studied 170 study words for 1.3 s. each. They were instructed to study the words for a later memory test. The study list consisted of 5 fillers at the beginning and end of the list. The 160 other study words consisted of 16 categories with category length 3 and 16 categories with category length 7. Half of categories were sampled from the pool of semantic categories and half were sampled from the pool of orthographic categories. The sampling was performed such that over all participants, each category from the pool was studied an approximately equal number of times. Half of the studied categories contained the prototype and half did not contain the prototype (an exemplar replaced the prototype). In the blocked condition, the categories were presented one after the other with the order of words within categories randomized as well the order of categories on the list. In the spaced condition, the order of all 170 study words (excluding the filler items) was randomized with the result that the category words were scattered over the study list. The Appendix B lists 9 exemplars per category. The study categories always contained the first two exemplars listed in Appendix B and never contained

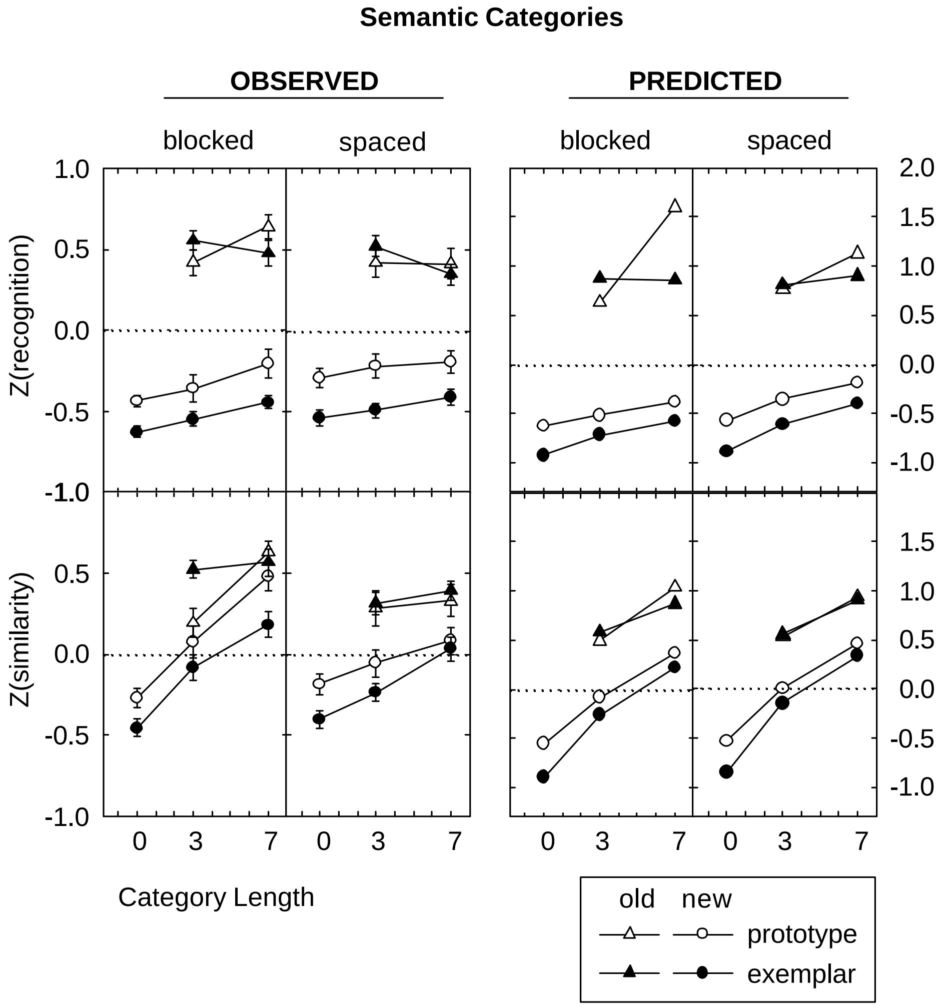


Figure 7. Observed and predicted results of Experiment 2, for the semantic categories.

the last two exemplars; they were reserved for testing as related distractors.

After the study list, participants were given instructions about the test phase. These instructions were identical to Experiment 1. Participants were given 112 test words for which they had to give recognition and similarity judgments as in Experiment 1. Table 3 lists the 18 conditions that were tested and the number of words that were tested per condition. For the testing phase, the exemplar words from the target conditions were always sampled from the first two exemplars from the list of Appendix B (these were also always sampled for the study list). The exemplar words for the distractor conditions of category length 3 and 7 were always drawn from the last two exemplars from the list.

There were two unrelated distractors conditions which we will refer to as prototype category length 0, and exemplar category length 0 conditions. In the first condition, the words were sampled from the 16 prototype words from the semantic and orthographic categories that were not studied (which categories were not studied varied from participant to participant). In the second category length 0 condition, the words were sampled from the last two exemplars of the 16 not studied semantic and orthographic categories. Because the same prototype or exemplar word could be tested as related distractors (category length 3 or 7) for participants that studied related words and as unrelated distractors (category length 0) for participants that did not study any words from that category, these conditions served as important controls for the related distractor conditions.

Results

The recognition and similarity judgments were converted to z-scores as in Experiment 1. The mean z-scores for the semantic and orthographic categories are shown in Figure 7 and 8 respectively. The ability of participants to discriminate between old and new items was computed with d' in the same manner as in Experiment 1. The d' results are listed in Table 1. Separate ANOVA's were performed on the target and distractor z-scores for the recognition and similarity ratings. Also, ANOVA's were performed on the sensitivity results on the recognition and similarity ratings. We will report the main effects of the within subject factors, category length (3 or 7) and category membership (exemplar or prototype) and the between subject factor, study presentation (blocked or spaced) and interactions between these factors. The differences in performance for the different category types (semantic or orthographic) will only be reported for the similarity ratings.

Recognition judgments. For targets, the main effects of category length and category membership were not significant [F(1,69)=1.50, MSE=.207, and F(1,69)=.056, MSE=.241, respectively]. However, Figures 7 and 8 show an interaction between category length and category membership. For category length 3, the confidence that the target is old was lower for prototype words than for exemplar words. However, for category length 7, the confidence that the target is old was higher for prototype words than for exemplar words. This interaction was significant

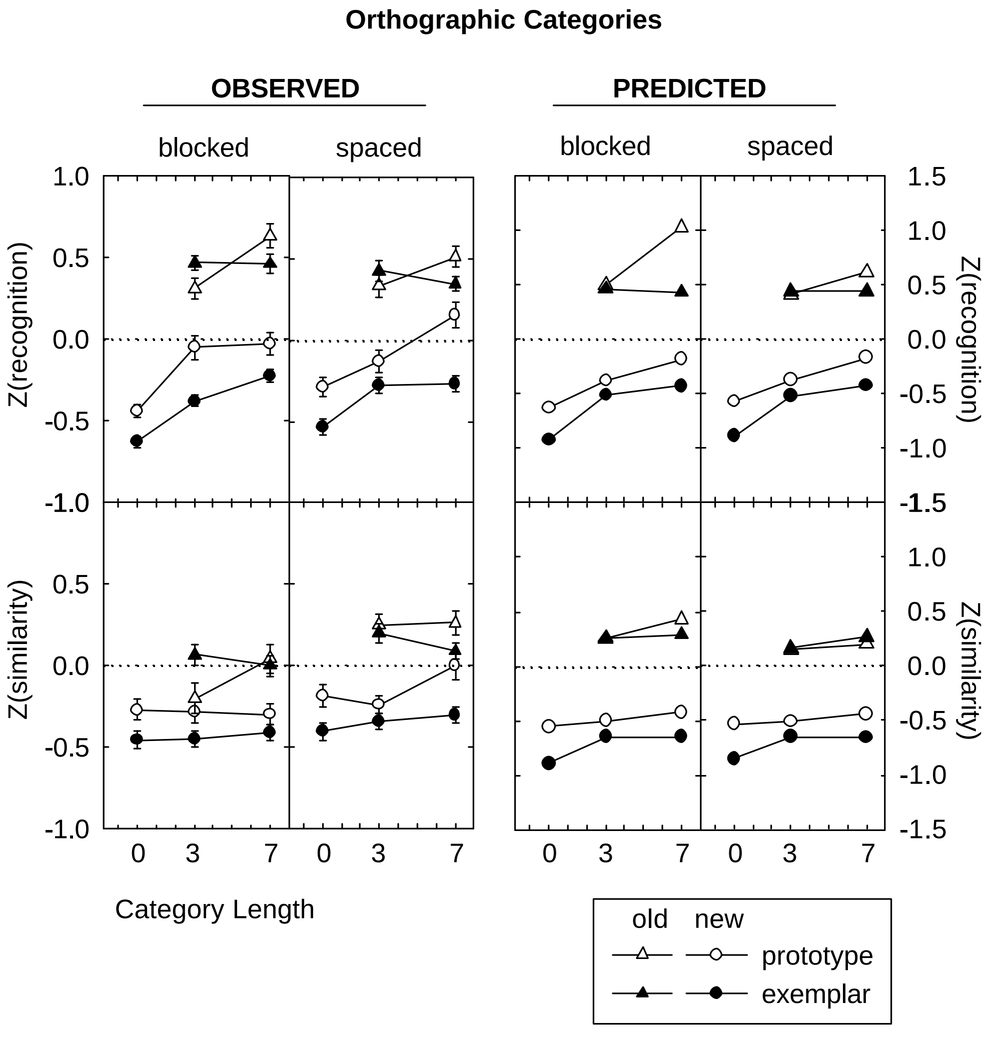


Figure 8. Observed and predicted results of Experiment 2, for the orthographic categories. Note that the two data points for category length 0 in the four panels for observed and predicted results are identical to the corresponding data points for the semantic categories shown in Figure 7.

[F(1,69)=17.84, MSE=.137].

For distractors, the confidence that the words were old increased with category length for both prototype and exemplar words. Also, the confidence that the distractor words were old was higher for prototype words than exemplar words for both category lengths. Both main effects of category length and category membership were significant [F(1,69)=10.6, MSE=.143, and F(1,69)=75.5, MSE=.119, respectively] while the interaction was not significant [F(1,69)=.364].

Table 1 shows that in the blocked condition, the ability to discriminate between old and new prototype words was higher for category length 7 than category length 3. For exemplar words, the pattern was reversed: the ability to discriminate between old and new exemplar words was lower for category length 7 than category length 3. This interaction between category length and category membership on sensitivity is significant [F(1,36)=12.20, MSE=1.13]. In the spaced condition, the effect of category length was the same on prototype and exemplar words: category length 7 old and new items were more difficult to discriminate than category length 3 old and new items. The interaction between category length and category membership was not significant [F(1,33)<1].

To simplify the analysis of the between subject factor of study presentation, three groups were created: targets, related distractors and unrelated distractors. The targets contained all target conditions, while the related distractor conditions contained all distractors with category length 3 or 7. The category length 0 distractors were pooled into the unrelated distractor group. Compared to spaced study presentation, blocked study presentation resulted in higher old ratings for targets and lower old ratings for both related and unrelated distractors. The average z-score ratings for blocked and spaced targets was .496 and .413 respectively, a significant difference [F(1,69)=8.25, MSE=.121]. For related distractors, the average z-score ratings for blocked and spaced study presentation was -.278 and -.230, a difference that did not reach statistical significance [F(1,69)=2.95, MSE=.112, p<.09] while for unrelated distractors, the average z-score ratings was -.531 and -.412, a significant difference [F(1,69)=5.91, MSE=0424].

The effect of study presentation on sensitivity was significant [F(1,69)=349, MSE=1.57]. As can be observed in Table 1, for most comparisons, the sensitivity was lower for the spaced study presentation than the blocked study presentation.

Similarity judgments. The pattern of results for the similarity judgments was similar to the pattern of results for the recognition judgments except for the effect of category length on distractors. For semantic categories, similarity ratings for distractors increased with category length. This effect was significant for both prototype and exemplar words [F(1,69)=44.6, MSE=.205, and F(1,69)=61.8, MSE=.164, respectively]. For orthographic categories, similarity ratings for distractors stayed more or less constant with category length. For orthographic categories, the effect of category length was not significant for either prototype or exemplar words [F(1,69)=1.28, MSE=.179, and F(1,69)=1.64, MSE=.117, respectively].

Number of ratings per word. In the spaced condition, a median of 6 participants (SD=2.07) rated each individual word. In the blocked condition, a median of 6 participants (SD=2.19) rated each individual word.

Discussion

There were several interesting patterns observed in the data. First, participants distinguished between the recognition and semantic similarity judgments. No effect of category length on semantic similarity judgments was observed for orthographic categories. This supports the assumption in the memory model that physical features such as orthographic features do not contribute in the generation of semantic similarity ratings. Second, effects of category length were observed for both semantic and orthographic categories which replicates the Shiffrin et al. (1995) results. Third, we did not fully replicate the differences between blocked and spaced study presentation as observed by Mather et al. (1997). We did replicate their observed effect of higher hit rates for targets and lower false alarm rates for unrelated distractors in the blocked vs. spaced condition. However, Mather et al. observed that false alarm rates for related distractors were higher in the blocked condition than in the spaced condition. We found a trend in the opposite direction. It is possible that the longer category length in the Mather et al. study explains this difference: their related distractors were related to more items in memory and perhaps the blocked presentation strongly evoked the false memory of the related distractor. The differences between blocked and spaced study presentation in this study suggest a recognition advantage for blocked over spaced presentation. This difference could be due to a variety of factors. For example, blocked presentation might lead to better memory organization that facilitates recognition judgments. In the modeling of these results, we will expand on one possible factor explaining these differences. It will be assumed that blocked presentation leads to stronger traces in memory (i.e., traces with more features). This could be because related words when blocked lead to better or more organized rehearsal leading in turn in stronger traces. Similarly, related words when blocked can activate each other implicitly leading to superior storage. More discussion on this assumption will follow in the modeling section.

As a last interesting aspect of the observed results, for targets, a crossover interaction was observed between category length and category membership. Target exemplars were better recognized than prototype exemplars when two related words were on the study list. However, target prototypes were better recognized than target exemplars when six related words were on the study list. This effect was observed for both orthographic and semantic categories and for both blocked and spaced study presentation. Based on this result only, it could be argued that this difference between prototypes and exemplars is due to differences in the process of storage or retrieval or both. However, the results for the distractors show that the difference between exemplars and prototypes is relatively constant when varying category length from 0 to 3 to 7. If it is assumed that the advantage of prototypes over exemplars for long category lengths was only due to retrieval differences, then an interaction between category length and category membership would be expected for distractors, a result that was not observed. Therefore, the results suggest that the cross-over interaction needs to be explained by differential storage advantages for prototypes in longer category lengths. It will be assumed that memory traces are especially strong for prototypes in the longer category lengths. Possible underlying mechanisms for this assumption are similar to the underlying mechanism for explaining the difference between blocked and spaced presentation. It is possible that the related exemplars implicitly activate the prototype word so that the presentation of the prototype word on the study list leads to strong traces in memory. Similarly, participants upon presentation of the prototype word could rehearse the prototype word more because the prototype word describes the semantic or orthographic category best. We will discuss this more in the modeling of these results.

Model Fits of Experiment 2

Two of the results from Experiment 2 require additions to the model applied to Experiment 1. 1) the difference between blocked and spaced study presentation (since the order of presentation was at first not assumed to play a role), and 2) the cross-over interaction between category length and category membership for targets. The second of these requires some discussion.

Note that the present model applied to Experiment 2 can predict differences between target prototypes and exemplars on the basis of word frequency differences. In Experiment 2 the prototype words had higher word frequency than the exemplar words, which could explain the lower hit rates for prototype than exemplar words for category length 3 (and higher false alarm rates for distractors). Also, the model can predict an interaction between category length and category membership for targets: prototype words are similar to more words than exemplars, and hence the log odds for these words grows faster than for exemplars as category length grows. However, some preliminary simulations suggested that the observed crossover interactions were too large for the model to predict adequately. Therefore, it was decided to augment the model to handle both this interaction and the blocked/spaced differences.

First, it is assumed that words in the blocked presentation condition lead to stronger traces in memory than in the spaced presentation condition. A justification relies on the possibility that participants notice the category structure, and such knowledge allows better rehearsal and coding. The probability of storing features in blocked and spaced words was parameterized by $u_{blocked}$ and u_{spaced}. These parameters were set at .8 and .7 respectively. Therefore, in the blocked condition, more complete traces were formed in memory than in the spaced condition. This predicts the observed result of higher hit rates and lower false alarm rates for blocked than spaced words. Second, it was assumed that there was a storage advantage for prototypes in the category length 7 condition and that this storage advantage was larger for blocked words than spaced words. A justification could be based on the development of IAR's for the prototype, IAR's that grow more prevalent as category length grows. Two parameters $u_{blocked,prot7}$ and $u_{spaced,prot7}$ were designated for the probability of storing features for the target prototypes of category length 7 in the blocked and spaced condition respectively. These were set respectively at 1.0 and .8 respectively. Together, the four parameters introduced in this section predict a storage advantage for blocked words over spaced words and prototype category length 7 words over all other words. One other change proved helpful in modeling in this study: the centering of responses for recognition and similarity judgments appeared different, so we allowed separate estimates of the centering parameter, γ: $\gamma_{recognition}$=.5, and $\gamma_{similarity}$=1.0.

Parameters. In addition to the parameters just discussed, there were the three basic parameters that were set at: c=.4, σ_n =.35, b=5.

Recognition and Similarity Judgments. The model's predictions for Experiment 2 are shown figure 7 and 8 for the semantic and orthographic categories respectively. The higher false alarm rates for prototype distractors over exemplar distractors was predicted by word frequency differences (the prototype words had higher word frequency than exemplar words). The cross-over interaction between category length and category membership for target items was predicted because of two factors. Because the prototype words had higher word frequency, a lower hit rate was predicted for the prototype words than exemplar words. However, because the prototype words for category length 7 are stored better than exemplar words, they are retrieved better. Together, these two factors combined to predict the cross-over interaction. For the semantic similarity judgments, the model predicted no category length effect for orthographic categories, because the orthographic features do not participate in the calculation of familiarity.

Sensitivity. Table 1 shows the predicted d' results. Overall, predicted d' was higher than observed. Since we only sought qualitative fits to the observed data, other parameter settings were not tried to lower the predicted d'. The pattern of predicted results for d' was similar to the pattern of observed results. Blocked presentation led to higher d' than spaced presentation. This was due to the stronger traces in the blocked presentation than spaced presentation.

Individual word correlations. The between subject factor of study presentation was collapsed for all correlational analyses of observed and predicted z-transformed ratings for individual words. This increased the median number of participants that rated each individual word to 12 (as opposed to 6 when the blocked and spaced conditions would be analyzed separately). Table 2 shows that the correlation between observed and predicted Z-scores for all words of Experiment 2 was .63 for the recognition judgments and .49 for the similarity judgments. When the scrambling procedure was applied, these correlations were reduced to .58 and .44 respectively. These reductions were statistically significant. This shows that most of the variance in performance was explained by between condition differences, including similarity factors, and that a small but significant portion was explained by similarity differences for individual words within condition.

Experiment 3

Some word pairs clearly have asymmetric associations between them. For example, the cue "fib" is strongly associated with "lie" but not vice versa. Ash and Ebenholtz (1962) have argued that the differences between forward and backward associations are not due to representational differences but because of process differences. If A->B is stronger than B->A, this is because the item B comes more readily to mind. Similarly, Nosofsky (1991) has argued that asymmetric similarities can be explained solely on the basis of stimulus differences such as strength, salience or frequency rather than on the basis of asymmetries underlying similarity relations.

In WAS, the similarity between word A and B is by definition equivalent to the similarity between words B and A. One way to predict asymmetries in performance utilizes word frequency differences. in the word association norms, it is almost invariably the case that if the association strength from A to B is stronger than from B to A (denoted by A->B), then the word frequency for A is lower than B. This is consistent with Ash and Ebenholtz (1962) and Nosofsky's (1991) view that the asymmetry can be explained by stimulus differences.

In this experiment, the idea is to use distractors that are forward, backward and bi-directional associatively related to target words and compare the performance for these related distractor words with unrelated distractor words that are either low or high frequency words. For example, suppose A is studied and F is tested as a distractor where F is a strong associate of A but not vice versa (i.e., A->F). Similarly, in other conditions, the false alarm rate of a word G is tested where G is backward associated to the studied word B but not vice versa (i.e., B<-G). The F words are almost guaranteed to be words with higher word frequencies than the G words. Based on these word frequency differences, a higher false alarm rate for the F words is predicted than for the G words. The interesting comparison is of the related distractor words F and G with unrelated distractor words with similar word frequencies. Differences between the false alarm rates for F and G and the unrelated distractor words that have similar word frequencies, cannot be due to word frequency and can only be explained on the basis of differences in semantic similarity. Specifically, the model predicts that the F and G words have higher false alarm rates than corresponding unrelated distractor conditions because the semantic features of the F and G words overlap more with the memory contents than unrelated distractor words.

Method

Design and participants. The design formed a (3 x 2) + 2 factorial design. The main factor was the directionality of association between study and test items and was varied in three levels: forward,

backward, and bi-directional. The second factor was oldness: words were tested as targets or distractors. The six conditions from these two factors were labeled A, B, C, F, G, and H. Words from the three target conditions A, B, and C, and three distractor conditions F, G, and H were drawn from associative pairs
A->F, B<-G, and C<->H respectively. Two distractor conditions were added with low and high frequency words that were unrelated to studied words. All conditions were tested in a within subject design. Sixty-two undergraduate students from the same pool of participants mentioned in Experiment 1 participated in the experiment.

Materials. Appendix C shows the words of this experiment. All words were selected from the pool of words from the production norms of Nelson et al. (1998). Two sets of 10 asymmetric associative word pairs, X->Y were created by selecting word pairs with strong forward and weak or absent backward associative strengths. The mean forward associative strength from X to Y was .812 (SD=.063) and mean backward associative strength from Y to X was .0301 (SD=.029). The mean Kucera and Francis frequency count was 2.05 (SD=2.31) for the X words and 76.8 (SD=72.3) for the Y words. One set of 10 bi-directional associative word pairs X<->Y was created by selecting word pairs with approximately equal forward and backward associative strengths. The mean forward and backward associative strengths was .356 (SD=.21). The mean Kucera and Francis frequency was 177 (SD=176) for these words. Two sets of 15 control words were created that were unrelated to the associatively related word pairs. The two sets contained low and high frequency words with mean frequencies of 2.00 (SD=1.13) and 306 (SD=106) respectively.

Procedure. Participants studied 120 study words for 1.3 s. each. They were instructed to study the words for a later memory test. The study list contained 90 filler words that were randomly selected from the pool of words from the production norms and 30 experimental words. These words contained an equal number of words from condition A, B, and C. Words from condition A were words with strong forward associations and weak backward associations (A->F). Words from condition B had the opposite pattern: weak forward associations and strong backward associations (B<-G). Words from condition C were words with strong forward and backward associations (C<->H). To control for word specific effects, two sets of words A, B, and C were created for the experiment. In set 1, the A, B, and C words were the left words of group 1, right words of group 2, and left words of group X<->Y words listed in Appendix C. In set 2, the A, B, and C words were the left words of group 2, right words of group 1, and right words of group X<->Y words listed in Appendix C. The participants were randomly assigned to one of two sets of experimental words. The order of the words on the study list was randomized for each participant with the constraint that 5 filler words were presented at the start and end of the study list.

After the study list, participants were given instructions about the test phase. These instructions were identical to Experiment 1. Participants were given 90 test words for which they had to give recognition and similarity judgments as in Experiment 1. The test words consisted of 30 old words and 60 new words. The 30 target words consisted of the 10 words from each the conditions A, B, and C. The 60 distractor words contained 30 distractors that were related to the study words and 30 words that were unrelated to the study words. The 30 related distractors consisted of 10 words from each of the conditions F, G, and H. Words from condition F were forward associatively related to the study words from condition A: they are produced as associates by A but do not produce A as associates (A->F). Words from condition G were backward associatively related to the study words of condition B (B<-G). Words from condition H were bi-directional associatively related to study words of condition H (C<->H). For participants who studied set 1 of experimental words, the words from conditions F, G, and H were selected from the right words of group 1, left words of group 2 and left words of group X<->Y from Appendix C. For participants who studied set 2 of experimental words, the words from conditions F, G, and H were selected from the left words of group 1, right words of group 2 and right words of group X<->Y from Appendix C. The 30 unrelated distractor words consisted of 15 low and 15 high frequency control words listed in Appendix C. The order of the test words was randomized for each participant.

Results and Discussion

As in Experiment 1 and 2, the recognition and similarity judgments were z-score transformed. The mean z-scores and standard errors for the three target, three related distractor, and two unrelated distractor conditions are shown in Figure 9. The d' results for several target-distractor condition comparisons are listed in Table 1. Separate ANOVA's were performed on the z-scores of target and distractor conditions. Also, ANOVA's were performed on the sensitivity results on the recognition and similarity ratings.

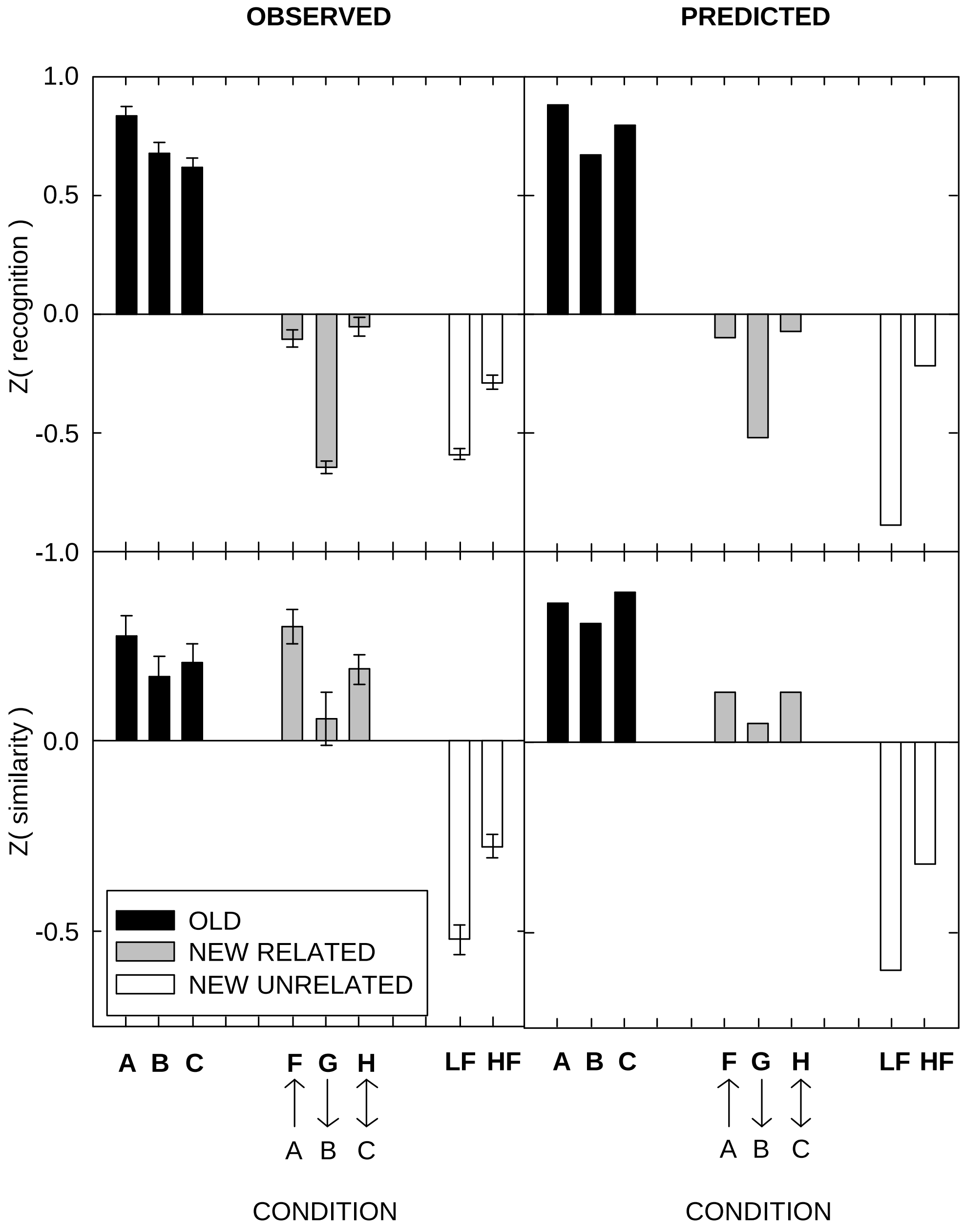


Figure 9. Observed and predicted results of Experiment 3.

Recognition judgments. Figure 9 shows that the target words from conditions A, B, and C were rated increasingly as less old. For the related distractor conditions, the lowest old ratings were given to words from condition G, while words from conditions F and H were given somewhat below average old ratings. The high frequency unrelated distractor words were rated significantly more old than the low frequency unrelated distractor words [F(1,61)=67.9, MSE=.0416]. The old ratings were significantly higher for A words than B words [F(1,61)=5.46, MSE=.136] while the old ratings were significantly higher for F words than G words [F(1,61)=146, MSE=.0626]. These differences are consistent with a mirror effect explanation based on word frequency differences. The B and F words were high frequency words while the A and G words were low frequency words: high frequency words tend to lead to lower hit and higher false alarm rates than low frequency words (i.e., the mirror effect, Glanzer & Adams 1985).

The interesting comparison is between unrelated and related distractor conditions that were similar in word frequency. The model predicted that old ratings should be higher for related distractors than unrelated distractors if the words have similar word frequencies. The high frequency words from related distractor conditions F and H were rated as significantly more old than the unrelated high frequency distractor words [F(1,61)=19.7, MSE=.053, and F(1,61)=20.2, MSE=.0828, respectively]. This confirms the prediction of the model. However, the unrelated low frequency distractor words were rated as more old than the words from condition G, a difference that did not reach statistical significance [F(1,61)=3.00, MSE=.0338, p<.088]. Because the model predicts that related distractors lead to higher old ratings than unrelated distractors, this observed trend in the opposite direction is an interesting finding.

Table 1 lists the participants' ability to discriminate between old and new words for various target and distractor conditions. The sensitivity in discriminating targets and distractors condition pairs was significantly lower for pairs that were forward associatively related (OLD-A vs. NEW-F) than pairs that were backward associatively related (OLD-B vs. NEW-G), [F(1,61)=30.9, MSE=.824].

Similarity judgments. The results for the similarity judgments were similar to the results of the recognition judgments with the difference that related distractors received similarity ratings that were about as high as the similarity ratings for target words. The d' results reflect that: the sensitivities of target and related distractor conditions are close to zero. Interestingly, the low frequency words from condition G that received lower recognition ratings than unrelated low frequency distractor words, received higher similarity ratings than the unrelated low frequency distractors [F(1,61)=66.1, MSE=.158].

Number of ratings per words. There were 41 and 21 participants that received study and test list 1 and 2 respectively. Since each participant rated all words from the pool of all possible test words, there were 41 and 21 ratings for each test word from sets 1 and 2 respectively.

Model Fits of Experiment 3

The model outlined in the Introduction, and applied to experiment 1, was applied to Experiment 3 without the special assumptions made for Experiment 2 .

Parameters. The four parameters to generate predictions for this experiment were set at: c=0.3, σ_n =.35, b=5, AND γ=3..

Recognition and Similarity Judgments. Figure 9 shows the predicted recognition and similarity results. In addition to the several ways in which the model made the correct predictions, there were some observed effects that were not handled well by the model. First, the difference between target conditions A and B was correctly predicted. The model predicted these differences based on word frequency. Words from condition A had lower word frequency than words from condition B. The difference between the unrelated low and high frequency distractors was also correctly predicted by word frequency differences. For the recognition ratings, the model predicted that related distractors from conditions F, G, and H have higher old ratings than the unrelated distractor conditions with similar word frequencies. This is because the related distractors overlap more with the memory contents than unrelated distractors. However, as was pointed out in the previous section, the results showed a trend for the condition G words to have lower old ratings than the unrelated low frequency distractors.

Another mismatch between observed and predicted results is for the condition C words. They were incorrectly predicted to have higher old ratings than condition B words despite the fact that the word frequency of condition C words was higher than condition B words. Also, the model incorrectly predicted that condition C words received the highest similarity ratings. This suggests that condition C words are not placed correctly with respect to the other study words (condition A and B) in the semantic space formed by WAS.

General Discussion

The memory model presented in this paper brings together the idea of explicit representation of orthographic and semantic features with a process model operating on those features. Words are represented by vectors of feature values that are based on an analysis of the semantic and orthographic features of words. The vectors of feature values representing various semantic aspects of words came from the Word Association Space. This space was developed by analyzing the associative relationships of a large database of free association norms and representing words with similar associative patterns with similar feature vectors. To represent orthography, the letters of the words were encoded. These representations were coupled with a process model for recognition memory. This model was based on the REM model, which used Bayesian principles to decide whether a memory probe is old or new.

One novel aspect in this model was the distinction between recognition and similarity judgments. The ability of participants to differentiate between recognition and similarity judgments was apparent in all experiments. Participants could distinguish between distractors that preserved the meaning of one of the themes on the study list versus distractors that were not similar to any words on the study list. In the model, the recognition judgments were assumed to rely on both the semantic and orthographic overlap of probe and memory contents while (semantic) similarity judgments were assumed to rely only on the semantic overlap of probe and memory contents. In Experiment 2, it was found that with orthographically related distractors, the category length of orthographic categories had no effect on (semantic) similarity judgments but increased the false alarms for the recognition judgments. This is consistent with the assumption that orthographic features are not involved in the calculation of similarity judgments.

The three experiments in this paper explored various predictions of the model with a focus on the interplay between semantic and orthographic similarity between probe and memory contents. The predictions of the model were tested at two different levels: at the level of condition means and at the level of individual word performance. In all three experiments, the model successfully predicted most of the qualitative differences in condition means. This suggests that the similarity relationships in the semantic space and in the orthographic representation are useful to predict memory performance.

Even stronger evidence for the idea that similarity relations among words explains recognition and similarity judgment data comes from the within condition correlation data. The correlational analyses showed that a small but significant part of the variance in performance was due to similarity relations due to differences among words within conditions, even though these words generally were chosen so such differences would be small.

An undesirable aspect of the present approach is the rather ad hoc fashion in which a word frequency mechanism had to be appended to the basic model. It may well be that a feature frequency approach would provide a more principled account, but this would only be possible in conjunction with a different word space, one in which high frequency words were clumped together, and one in which high frequency words had high frequency features that were clumped near the center of each featural dimension. **WAS** represented similarity by inner products, resulting in high frequency words being pushed to the outside of the space. This problem was solved by normalizing the vector lengths, but at the cost of removing word frequency differences, in turn requiring the model to be augmented by a different word frequency mechanism of a very ad hoc nature. There is obviously room here for further research and improvement of the models.

There are several areas ways in which the model can be extended and there are several new assumptions that can be tested. For example, one major assumption in the REM model and this memory model is that the features that represent different aspects of words can be stored in one trace. Instead, it could be assumed that separate attributes such as semantic and physical features are stored in separate traces. This would lead to a system in which familiarity is calculated for the semantic and physical contents of memory separately as opposed integrally. Preliminary simulations have suggested that there is not much difference between these two recognition memory models.

Notes

1. In part I of this research, Table 5, it was shown that with and without the normalization of the vector lengths, WAS is sensitive to semantic information because it predicts much larger within category similarities than between category similarity where the categories were defined semantically.

2. The distribution g of all stored feature values was determined by integrating over the probe feature distribution and noise distribution: each stored feature value could have been produced by a combination of each probe feature value and some noise value.

3. Even though the model is quite simple in its mathematical form, the calculations are computationally very involved because of Equation (5), in which the likelihood ratio is calculated for a single trace with 400 semantic features. The computational requirements of the simulations prevented us from applying model fitting procedures.

4. Because the scrambling is random, the obtained correlation obtained with the scrambling procedure is itself a stochastic variable. We report the correlation that is an average of the correlation by performing scrambling procedure 100 times.

5. The u storage parameter that determined the probability that orthographic and semantic features were stored was set at one so that this part of this storage process that determines the strength of traces in memory was effectively not used.

References

Agostino, P.R. (1969). The blocked-random effect in recall and recognition. Journal of Verbal Learning and Verbal Behavior, 8, 815-820.

Anisfeld, M., & Knapp, M. (1968). Association, synonymity, and directionality in false recognition. Journal of Experimental Psychology, 77, 171-179.

Ash, S.E. & Ebenholtz, S.M. (1962). The principle of associative symmetry. Proceedings of the American Philosophical Society, 106, 135-163.

Baddeley, A.D., & Dale, H.C.A. (1966). The effect of semantic similarity on retroactive interference in long- and short-term memory. Journal of Verbal Learning and Verbal Behavior, 5, 417-420.

Bower, G.H. (1967). A multicomponent theory of the memory trace. In K.W. Spence & J.T. Spence (Eds.), The psychology of learning and motivation, Vol 1. New York: Academic Press.

Brainerd, C.J., Reyna, V.F., & Mojardin, A.H. (1999). Conjoint recognition. Psychological Review, 106, 160-179.

Brainerd, C.J., & Reyna, V.F. (1998). When things that were never experienced are easier to “remember” than things that were. Psychological Science, 9, 484-489.

Bransford, J.D., & Franks, J.J. (1972). The abstraction of linguistic ideas: a review. Cognition, 1, 211-249.

Bregman, A.S. (1968). Forgetting curves with semantic, phonetic, graphic, and contiguity cues. Journal of Experimental Psychology, 78, 539-546.

Brown, R., & McNeil, D. (1966). The “tip of the tongue” phenomenon. Journal of Verbal Learning and Verbal Behavior, 5, 325-337.

Burnage, D. (1998). CELEX: a guide for users. Nijmegen, Netherlands: Centre for Lexical Information.

Buschke, H., & Lenon, R. (1969). Encoding homophones and synonyms for verbal discrimination and recognition. Psychonomic Science, 14, 269-270.

Canas, J. J. (1990). Associative strength effects in the lexical decision task. The Quarterly Journal of Experimental Psychology, 42, 121-145.

Cermak, G., Schnorr, J., Buschke, H., & Atkinson, R.C. (1970). Recognition memory as influenced by differential attention to semantic and acoustic properties of words. Psychonomic Science, 19, 79-81.

Cramer, P. (1968).Word Association. NY: Academic Press.

Davies, G., & Cubbage, A. (1976). Attribute coding at different levels of processing. Quarterly Journal of Experimental Psychology, 28, 653-660.

Deese, J. (1959a). Influence of inter-item associative strength upon immediate free recall. Psychological Reports, 5, 305-312.

Deese, J. (1959b). On the prediction of occurrences of particular verbal intrusions in immediate recall. Journal of Experimental Psychology, 58, 17-22.

Deese, J. (1962). On the structure of associative meaning. Psychological Review, 69, 161-175.

Deese, J. (1965). The structure of associations in language and thought. Baltimore, MD: The Johns Hopkins Press.

Dennis, S. (1995). The Sydney Morning Herald Word Database. Noetica: Open Forum, 1(4), http://psy.uq.edu.au/CogPsych/Noetica/.

Dennis, S. & Humphreys, M.S. (1998). Cueing for context: An alternative to global matching models of recognition memory. In M. Oaksford & N. Chater (Eds.). Rational models of cognition. (pp. 109-127), Oxford, England: Oxford University Press.

Dennis, S., & Humphreys, M.S. (submitted). A context noise model of episodic word recognition.

Diller, D. E., Nobel, P. A., and Shiffrin, R. M. (in press). An ARC-REM model for accuracy and response time in recognition and recall. Journal of Experimental Psychology: Learning, Memory, and Cognition.

Eich, J.M. (1982). A composite holographic associative recall model. Psychological Review, 89, 627-661.

Elias, C.S., & Perfetti, C.A. (1973). Encoding task and recognition memory: the importance of semantic encoding. Journal of Experimental Psychology, 99, 151-156.

Gardiner, J.M., & Java, R.I. (1991). Forgetting in recognition memory with and without recollective experience. Memory & Cognition, 19, 617-623.

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

Glanzer, M., & Adams, J.K. (1990). The mirror effect in recognition memory: data and theory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 16, 5-16.

Glanzer, M., & Bowles, N. (1976). Analysis of the word-frequency effect in recognition memory. Journal of Experimental Psychology: Human Learning and Memory, 2 (1), 21-31.

Gorman, A. N. (1961). Recognition memory for names as a function of abstractness and frequency. Journal of Experimental Psychology, 61, 23-29.

Hall, J.W. & Kozloff, E.E. (1973). False recognition of associates of converging versus repeated words. American Journal of Psychology, 86, 133-139.

Herriot, P. (1974). Attributes of memory. London: Methuen.

Hintzman, D.L. (1984). Minerva 2: a simulation model of human memory. Behavior Research Methods, Instruments, and Computers, 16, 96-101.

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

Huber, D. E., Shiffrin, R. M., Lyle, K. B., & Ruys, K. I. (in press). Perception and preference in short-term word priming. Psychological Review

Kinsbourne, M., & George, J. (1974). The mechanism of the word frequency effect on recognition memory. Journal of Verbal Learning and Verbal Behavior, 13, 63-69.

Kim, K. & Glanzer, M. (1993). Speed versus accuracy instructions, study time, and the mirror effect. Journal of Experimental Psychology, Learning, Memory, and Cognition, 19, 638-652.

Krumhansl, C.L. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. Psychological Review, 85, 445, 463.

Kucera, H., & Francis, W.N. (1967). Computational analysis of present-day American English. Providence, RI: Brown University Press.

Landauer, T.K., & Dumais, S.T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.

Landauer, T.K., & Streeter, L.A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. Journal of Verbal Learning and Verbal Behavior, 12, 119-131.

Laurence, M.W. (1970). Role of homophones in transfer learning. Journal of Experimental Psychology, 86, 1-7.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. Ear and Hearing, 19, 1-36.

Mandler, G. (1980). Recognizing: the judgment of previous occurrence. Psychological Review, 87, 252-271.

Mather, M., Henkel, L.A., & Johnson, M.K. (1997). Evaluating characteristics of false memories: remember/know judgments and memory characteristics questionnaire compared. Memory & Cognition, 25, 826-837.

McClelland, J.L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. Psychological Review, 105, 724-760.

McCormack, P. D., & Swenson, A. L. (1972). Recognition memory for common and rare words. Journal of Experimental Psychology, 95, 72-77.

Miller, M.B., & Wolford, G.L. (1999). Theoretical commentary: the role of criterion shift in false memory. Psychological Review, 106, 398-405.

Morton, J.A. (1970). A functional model for memory. In D.A. Norman (Ed.), Models of human memory. New York: Academic Press.

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

Nelson, D.L. (1972). Words as sets of features: The role of phonological attributes. In R.F. Thompson & J.F. Voss (Eds.), Topics in learning and Performance. New York: Academic.

Nelson, D.L., & Brooks, D.H. (1973). Independence of phonetic and imaginal features. Journal of Experimental Psychology, 97, 1-7.

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

Nelson, D. L., Schreiber, T. A., & McEvoy, C. L. (1992). Processing implicit and explicit representations. Psychological Review, 99, 322-348.

Nosofsky, R.M. (1991). Stimulus bias, asymmetric similarity, and classification. Cognitive Psychology, 23, 94-140.

Norman, D.A., & Rumelhart, D.E. (1970). A system for perception and memory. In D.A. Norman (Ed.), Models of human memory. New York: Academic Press.

Payne, D.G., Elie, C.J., Blackwell, J.M., & Neuschatz, J.S. (1996). Memory illusions: Recalling, and recollecting events that never occurred. Journal of Memory and Language, 35, 261-285.

Pike, R. (1984). Comparison of convolution and matrix distributed memory systems for associative recall and recognition. Psychological Review, 91, 281-293.

Robinson, K.J., & Roediger, H.L., III, (1997). Associative processes in false recall and false recognition. Psychological Science, 8, 231-237.

Roediger, H.L., & McDermott, K.B. (1995). Creating false memories: remembering words not presented on lists. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21, 803-814.

Runquist, W.N., Blackmore, M. (1973). Phonemic storage of concrete and abstract words with auditory presentation. Canadian Journal of Psychology, 27, 456-463.

Scarborough, D., Cortese, C., & Scarborough, H. (1977). Frequency and repetition effects in lexical memory. Journal of Experimental Psychology: Human Perception and Performance, 3, 1-17.

Schacter, D.L., Verfaellie, M., & Pradere, D. (1996). The neuropsychology of memory illusions: false recall and recognition in amnesic patients. Journal of Memory & Language, 35, 319-334.

Schooler, L .J., Shiffrin, R. M., & Raaijmakers, J. G. W. (in press). Theoretical note A Bayesian model for implicit effects in perceptual identification. Psychological Review.

Shepard, R.N. (1967). Recognition memory for words, sentences, and pictures. Journal of Verbal Learning and Verbal Behavior, 6, 156-163.

Shiffrin, R.M., Huber, D.E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21, 267-287.

Shiffrin, R.M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. Psychonomic Bulletin & Review, 4, 145-166.

Shiffrin, R. M., & Steyvers, M. (1998). The effectiveness of retrieval from memory. In M. Oaksford & N. Chater (Eds.). Rational models of cognition. (pp. 73-95), Oxford, England: Oxford University Press.

Sommers, M.S., & Lewis, B.P. (1999). Who really lives next door: creating false memories with phonological neighbors. Journal of Memory and Language, 40, 83-108.

Toglia, M.P., Hinman, P.J., Dayton, B.S., & Catalano, J.F. (1997). The blocked-random effect in pictures and words. Perceptual and Motor Skills, 84, 976-978.

Tussing, A.A., & Greene, R.L. (1997). False recognition of associates: How robust is the effect? Psychonomic Bulletin & Review, 4, 572-576.

Underwood, B.J. (1965). False recognition produced by implicit verbal responses. Journal of Experimental Psychology, 70, 122-129.

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

Watkins, M.J., Watkins, O.C., & Crowder, R.G. (1974). The modality effect in serial and free recall as a function of phonological similarity. Journal of Verbal Learning and Verbal Behavior, 13, 430-447.

Wickens, D.D. (1972). Characteristics of word encoding. In A.W. Melton & E. Martin (Eds.), Coding processes in human memory. Washington, D.C.: V.H. Winston, pp. 191-215.

Wickens, D.D., Ory, N.E., & Graf, S.A. (1970). Encoding by taxonomic and acoustic categories in long-term memory. Journal of Experimental Psychology, 84, 462-469.

Zechmeister, E. B. (1969). Orthographic distinctiveness. Journal of Verbal Learning and Verbal Behavior, 8, 754-761.

Appendix A
Words of Experiment 1

#	WF	Study Words	Test Words
1	L	RATTLE, REPTILE, VENOM, COBRA	SERPENT, LIZARD, FANGS, PONY
	H	BITE, WORM, GRASS, POISON	SNAKE, DEATH, SLIMY, HEAL
2	L	ROYAL, PRINCE, PALACE, CHESS	THRONE, EMPEROR, GROOM, RUBY
	H	CASTLE, KING, RULER, PRINCESS	QUEEN, CROWN, LEADER, FANTASY
3	L	OAR, ROW, VESSEL, SAILING	YACHT, RAFT, CANAL, REFLECT
	H	CAPTAIN, SHIP, SAIL, BOAT	SAILOR, NAVY, RIVER, HEAT
4	L	UNTRUTHFUL, FIB, DECEPTION, RUMOR	PERJURY, FRAUD, SINCERE, IMPRESSION
	H	FALSE, CHEAT, TRUE, TRUTH	DENY, LIAR, FACT, SHORT
5	L	GARAGE, BUMPER, DRIVEWAY, AUTOMOBILE	VAN, WINDSHIELD, COMPACT, LEVER
	H	TRUCK, DRIVE, DRIVER, TIRE	VEHICLE, WHEEL, BUS, ICE
6	L	BLAST, ERUPT, BURST, ATOMIC	EXPLOSION, DYNAMITE, NOISY, SHIVER
	H	BOMB, BLOW, BANG, NOISE	LOUD, BOOM, SOUND, POOL
7	L	MUFFIN, STALE, ROLL, CRUST	BISCUIT, BAKER, SLICE, DIVER
	H	BUTTER, WHEAT, BREAD, TOAST	DOUGH, JELLY, CAKE, WEAK
8	L	DUNGEON, CAPTIVE, CELL, PROSECUTE	CONVICT, INMATE, FUGITIVE, DISGRACE

	H	PRISON, CRIMINAL, PUNISHMENT, PRISONER	JAIL, CRIME, COURT, ANGRY
9	L	STEEPLE, BAPTIST, MINISTER, PRAYER	CATHEDRAL, SYNAGOGUE, BLESSING, PHILOSOPHY
	H	PRIEST, CATHOLIC, RELIGION, TEMPLE	CHURCH, BIBLE, FAITH, HAT
10	L	PETALS, DAISY, STEM, TULIP	BLOOM, VIOLET, MEADOW, COCOON
	H	GARDEN, ROSE, FLOWER, VASE	PLANT, SEED, POT, LADY
11	L	FLASH, BOLT, VOLT, UMBRELLA	THUNDER, BEAM, FLASHLIGHT, DELAY
	H	CLOUD, BRIGHT, STORM, ELECTRICITY	RAIN, WIND, SNOW, PENCIL
12	L	BIZARRE, UNCOMMON, ABNORMAL, ORDINARY	INSANE, IRREGULAR, AWKWARD, TWICE
	H	WEIRD, UNIQUE, CRAZY, COMMON	STRANGE, AVERAGE, WILD, KIND
13	L	ADORE, AFFECTION, CUDDLE, SWEETHEART	PASSION, VALENTINE, AFFAIR, SNOTTY
	H	CARE, LIKE, ROMANCE, MOUTH	KISS, RELATIONSHIP, MARRIAGE, ROUGH
14	L	MEEK, BASHFUL, TIMID, INTROVERT	MODEST, WITHDRAWN, HUMILIATE, CHIME
	H	SILENT, QUIET, EMBARRASS, OUTGOING	SHY, CALM, SECRET, VACATION
15	L	INSTRUCTOR, LEARNER, INSTRUCT, PUPIL	EDUCATE, FACULTY, CHALKBOARD, ESTEEM
	H	INTELLIGENT, LEARN, STUDENT, TEACH	PROFESSOR, COLLEGE, BRAIN, CLOCK
16	L	SKUNK, SNIFF, FRAGRANCE, COLOGNE	STENCH, FOUL, CIGAR, THANKSGIVING
	H	TASTE, STINK, NOSE, PERFUME	ODOR, SENSE, ONION, MALE
17	L	GIGGLE, PRANK, RIDDLE, COMEDIAN	HILARIOUS, AMUSE, WIT, FRECKLE
	H	JOKE, LAUGH, CLOWN, CRY	FUNNY, COMEDY, SMILE, ANXIOUS
18	L	WEAVE, SEAM, KNIT, CROCHET	STITCH, SPOOL, PRICK, TREND
	H	STRING, SEW, PIN, THREAD	NEEDLE, YARN, CLOTH, SING

Notes: WF=word frequency; the test words appear in order of semantic similarity according to the semantic space

Appendix B
Words of Experiment 2

Prototypes	Exemplars
Semantic Categories	
LOAN	CASH, FEE, FUND, BANKER, OWE, DEBT, CHECKBOOK, PROFIT, DEPOSIT
BELIEVE	DECEIVE, RUMOR, TRUTH, DECEPTION, FACT, LIAR, FIB, HONESTY, BLUFF
DOOR	ENTRANCE, KNOB, HALLWAY, KEY, LOCK, CORRIDOR, MAT, HINGE, THRESHOLD
DUST	GRIT, FILTH, SOOT, SCUM, GROUND, DIRT, PILE, SOIL, MUD
WET	MOIST, RAINY, DRENCH, DEW, GALOSHES, SLIPPERY, SOAK, PUDDLE, DAMP
KING	CROWN, THRONE, EMPEROR, MONARCH, CASTLE, PALACE, PRINCESS, ROYALTY, QUEEN
AFRAID	PANIC, FRIGHT, TERROR, SUPERSTITION, FEAR, MONSTER, HAUNT, SPOOK, SCARED
VICTORY	CONQUER, DEFEAT, CONTEST, COMPETE, CHAMPION, AWARD, TRIUMPH, TROPHY, WIN
JUDGE	LAWYER, VERDICT, ATTORNEY, WITNESS, COURT, TESTIFY, EVIDENCE, GAVEL, TRIAL
TRAIN	UNDERGROUND, CONDUCTOR, CABOOSE, SUBWAY, EXPRESS, TUNNEL, WAGON, CROSSING, STATION
HUSBAND	COMPANION, ENGAGE, PARTNER, FAITHFUL, MATE, LOVER, WED, SPOUSE, MARRY
PHONE	BOOTH, CORD, DIAL, COMMUNICATION, OPERATOR, SPEAKER, EXTENSION, MESSAGE, RUNG
WINTER	SHIVER, FRIGID, FROST, IGLOO, ICEBERG, CHILL, ARCTIC, FREEZER, COLD
SLEEP	SNOOZE, REST, HAMMOCK, WAKE, PAJAMAS, SLUMBER, DROWSY, NAP, NIGHTGOWN
EYE	CONTACTS, VISION, FOCUS, SQUINT, SEE, LENS, VIEW, BLIND, LASH

BLOOD	PLASMA, DONOR, FLESH, ARTERY, VAMPIRE, PRICK, DRACULA, TRANSPLANT, VEIN
POLITICS	CANDIDATE, LEGISLATURE, DEMOCRACY, CONGRESS, LEADERSHIP, PRESIDENT, CAMPAIGN, GOVERNMENT, SENATOR
SUIT	BUTTON, VEST, COLLAR, BLOUSE, SHIRT, TUXEDO, JACKET, LAPEL, KNOT
HORSE	SADDLE, TROT, UNICORN, COLT, MARE, RANCH, STABLE, RODEO, GALLOP
CHAIR	BENCH, SEAT, TABLE, WICKER, STOOL, COUCH, SOFA, RECLINER, SITTING
TELEVISION	PROGRAM, CHANNEL, ANTENNA, NETWORK, ENTERTAINMENT, ADVERTISEMENT, CABLE, MEDIA, COMMERCIAL
DINNER	CHINESE, FEAST, BANQUET, THANKSGIVING, MEAL, CAFETERIA, SUPPER, TRAY, LUNCH
SNAKE	SERPENT, RATTLE, DEADLY, SLITHER, COBRA, BITE, LIZARD, VENOM, REPTILE
TEXT	ALMANAC, AUTHOR, LITERATURE, PAGE, PUBLISHER, LIBRARY, READER, NOVEL, CHAPTER

Orthographic Categories

BAG	BEG, BAN, BAR, BAT, BUG, BAD, BIG, BAY, BOG
HOT	HIT, HUT, HOE, HOG, HOW, LOT, HAT, HOP, DOT
RAW	RAY, JAW, RAP, RAT, RAM, ROW, PAW, LAW, RAG
SIN	SIX, GIN, SIP, SUN, SON, FIN, SIT, PIN, KIN
DIE	ACE, TIE, PIE, DIM, LIE, DIG, DUE, DIP, DOE
TEN	TAN, BED, TON, PEN, TIN, HEN, TEA, MEN, BEE
BEAT	BENT, BOAT, BEST, BEAR, BELT, BEAD, BEET, BEAN, BEAM
CARE	CASE, CAGE, CAPE, CANE, CART, CAKE, CARD, CAFE, CAVE
LINE	DINE, LIKE, LICE, LINT, LIFE, LANE, LINK, LIME, FINE
HALL	HALF, HAUL, FALL, CALL, HALT, BALL, HAIL, HALO, HELL
MASS	BASS, MESS, MARS, BASE, MISS, MASH, PASS, MOSS, MASK
FORM	FORK, DORM, FIRM, WORM, FOAM, FORT, BOOM, NORM, FARM
LEAD	DEAD, LEAP, HEAD, LEAN, READ, LEAK, LEAF, LOAD, LEND
SALE	SAFE, TALE, SALT, SAVE, SAGE, MALE, SAME, PALE, SOLE
RACE	RARE, BAKE, RAGE, RACK, LACE, RICE, RAKE, RATE, FACE
WIDE	WIRE, TIDE, WINE, WIPE, HIDE, RIDE, WISE, WIFE, SIDE
FILL	FILM, FELL, BILL, KILL, FULL, MILL, FILE, DILL, HILL
LOST	LAST, LOSS, COST, LOSE, MOST, LUST, LOFT, HOST, LIST
SHARE	SHAPE, SCARE, SHAKE, SHAME, SHAVE, SHADE, SHARK, SHORE, SHARP
GRACE	BRAKE, TRACE, BRAVE, GRADE, GRAVE, CRACK, GRAZE, GRAPE, GRATE
FIGHT	RIGHT, SIGHT, EIGHT, DIGIT, MIGHT, FIRST, NIGHT, TIGHT, LIGHT
MATCH	DITCH, LATCH, HATCH, MARSH, MARCH, WATCH, CATCH, PATCH, HITCH
PRIME	PRIDE, PRICE, BRIBE, CHIME, PRIZE, GRIME, CRIME, BRIDE, DRIVE
ROUND	SOUND, COUNT, ROUGH, ROUGE, HOUND, FOUND, BOUND, WOUND, POUND

Note: the last two words of each category are only tested as new words

Appendix C
Words of Experiment 3

X->Y (Group 1)	X->Y (Group 2)	X<->Y
FIB - LIE	KIWI - FRUIT	PRIVATE - PUBLIC
MOO - COW	SWATTER - FLY	ACTION - REACTION
MEOW - CAT	DASHBOARD - CAR	CAUSE - EFFECT
TARDY - LATE	SCISSORS - CUT	ALONE - LONELY
GLACIER - ICE	TROUT - FISH	FOOD - EAT
GIGGLE - LAUGH	SLIPPERY - WET	GIRLS - BOYS
HILARIOUS - FUNNY	BLAZE - FIRE	GOOD - BAD
BOUQUET - FLOWERS	BRAWL - FIGHT	ADMIRE - RESPECT
TELLER - BANK	BUMBLE - BEE	DECISION - CHOICE
DESPISE - HATE	CHIRP - BIRD	SAD - HAPPY

Low Frequency Control	High Frequency Control
SAXOPHONE	WIFE
ABUSE	THING
CROCHET	SHORT
GRANITE	COMPANY
SKYSCRAPER	TODAY
LOSER	PROGRAM
BURGLARY	EVIDENCE
HANDCUFFS	GENERAL
SURF	LAND
CAULIFLOWER	SOUND
LATHER	ART
ASHTRAY	COURSE
CONCEIT	EYES
CLENCH	FORCE
INSTRUCT	THOUGHT

Part III: Feature Frequency Effects in Recognition Memory

Low frequency words are better recognized than high frequency words (Glanzer & Adams, 1985; McCormack & Swenson, 1972; Schulman, 1967; Shepard, 1967; but see Wixted, 1992), a phenomenon known as the word-frequency effect. For single-item yes-no recognition (i.e. old-new), hit rates (correctly responding "old" to an old item) are higher for low frequency words than for high frequency words and false alarm rates (incorrectly responding "old" to a new item) are higher for high frequency words than low frequency words (McCormack & Swenson, 1972; Glanzer & Adams 1985; Schulman, 1967; Shepard, 1967).

Several different explanations for the word-frequency effect have been proposed; probably because word frequency is correlated with many variables. The advantage for low frequency words has been attributed to elevated attention (Brown, 1976; Glanzer & Adams, 1990; Lockhart, Craik, & Jacoby, 1976; Maddox & Estes, 1997; Shepard, 1967), extra rehearsal time (Mandler, 1980), differences in pre-experimental recency (Scarborough, Cortese, & Scarborough, 1977; Underwood & Schultz, 1960), noise from extra-list memory (Estes, 1994; Maddox & Estes, 1997; Shiffrin & Steyvers, 1997), number of different contexts (Dennis & Humphreys, in review) and differences in the variability with which words are encoded (McClelland & Chappell, 1998). The Retrieving Effectively from Memory theory (REM, Shiffrin & Steyvers, 1997, 1998) accounts for the word-frequency effect on the assumption that the memory representations of low-frequency words tend to be made up of less common features than the memory representations of high-frequency words. It is of course possible that several or all of the mechanisms proposed are operating simultaneously. It should be pointed out that while Shiffrin and Steyvers (1997) employed the feature frequency assumption as the sole mechanism to predict word frequency effects, they were careful to point out that many other plausible factors could also contribute to word frequency effects. In this paper, however, we empirically test the feature-frequency assumption.

Landauer and Streeter (1973) pointed out that the frequency distributions of orthographic and phonetic features are dependent on normative word-frequency. For example, the letter "X" is twice as likely to occur in rare words than in common words. Almost all implementations of the REM model assume that features vary in their environmental frequency, or 'base rate'. This feature frequency assumption can be used to explain word frequency effects: because high frequency words are encountered more often, the features that make up high frequency words are also encountered more often. This means that feature frequency is correlated with normative word frequency. In REM (Shiffrin & Steyvers, 1997), high frequency words were represented with vectors having more common feature values and low frequency words were represented with vectors having more rare feature values. Because the REM model is sensitive to the diagnosticity of the features that make up words (memory traces with rare features that match the test features provide better evidence), it predicted an advantage for low frequency words over high frequency words as well as mirror effects for hit and false alarm rates.

Convergent evidence for the feature-frequency assumption comes from a set of experiments by Zechmeister (1969, 1972) that showed that words that were rated as orthographically distinct (e.g. sylph) were better recognized than words rated less orthographically distinct (e.g. parse). He also showed that the distinctiveness ratings were related to both the frequency of letter combinations and orthographic distinctiveness.

In this study, instead of using ratings, we assess feature frequency by measures that are directly based on the frequencies of the individual letters that make up words. The results of this study will be modeled by two versions of the REM model. The first model is based on the REM model as described by Shiffrin and Steyvers (1997) in which words are represented by arbitrary feature values. In the second model, the representation of the words is directly based on the orthography of the words used in the experiment and on the environmental base rates of letters occurring in words.

Experiment

Feature frequency and natural language word frequency are correlated variables: the frequency of a word determines the frequency of the letters that occur in the word. The experiment was designed to test the hypothesis that the frequency of occurrence of orthographic features in natural language, operationally defined as letters, affects the recognition of words independently of natural language word frequency. According to the feature-frequency account of the word-frequency effect for recognition (Shiffrin & Steyvers, 1997; Zechmeister, 1969, 1972), words comprised primarily of low-frequency letters should be better recognized than words comprised primarily of high-frequency letters, independent of other factors correlated with a word's normative frequency. In contrast, if orthographic feature frequency does not affect word recognition, then words comprised of common letters and words comprised of uncommon letters should be recognized equally well, if both groups are of equal normative word frequency.

Method

Participants. Fifty-three Indiana University students who were enrolled in introductory psychology courses participated in exchange for course credit.

Design and Materials. Normative word frequency and normative letter frequency were manipulated as within-subject factors in a 2 x 2 factorial design. The dependent variables were the probability of responding "old" and sensitivity operationally defined as $\underline{d}_a$ (Macmillan & Creelman, 1991; Swets & Pickett, 1982).

Two hundred and eighty-eight words were selected from the CELEX database (Burnage, 1998). The stimuli were organized into four groups (72 in each), according to orthographic feature frequency and normative word frequency: low feature frequency, low word frequency (LFF-LWF); high feature frequency, low word frequency (HFF-LWF); low feature frequency, high word frequency (LFF-HWF); and high feature frequency, high word frequency (HFF-HWF). The stimuli are listed in the Appendix A1.

High-frequency words were operationally defined as those occurring between 15 and 39 times per million of words in the natural language and low-frequency words were as those occurring between 3 and 7 times per million of words in the natural language. Orthographic feature frequency was operationally defined in the following manner. The relative frequencies of letters occurring in the first, interior, and the final positions of the words included in the CELEX database were computed as follows: in each of these three positions, if a letter was found in a word it was counted as having occurred as many times as the frequency count of that word in the language (per million). Thus each letter was weighted by the normative frequencies of the words in which a letter appeared. Table 1 lists the resultant orthographic feature frequencies of the first, interior, and final positions. Note for example that the letter "y" is the fourth most frequent letter at the ending of a word but is the fifth least frequent letter in the interior positions of a word.

The overall orthographic feature frequency of a given word was then measured in two different ways. In the first measure (referred to as feature frequency A), for each word, the product was calculated of the relative letter frequencies of the letters in their corresponding positions in the word. For example, using Table 1, the word "bane" would get a measure of (.0476)(.1157)(.0578)(.2592) = 0.000082 and the word "ajar" would get a measure of (.0556)(.00078)(.1157)(.0933) = 0.0000047. In a second measure (referred to as feature frequency B), the average relative letter-frequencies of the letters in their corresponding positions was calculated. According to this measure, the words "bane" and "ajar" would get measures of ((.0476)+(.1157)+(.0578)+(.2592))/4 = .12 and ((.0556)+(.00078)+(.1157)+(.0933))/4 = .066 respectively. According to both measures A and B, the word "bane" consists of more high frequency letters than the word "ajar". The words "bane" and "ajar" are examples of words in the HFF-LWF and LFF-LWF respectively since the words differ in their feature frequencies (by measures A and B) and both words have low word frequency (3 per million).

Table 1
Relative frequencies of letters in first, interior and last word positions

Rank	First		Interior		Last	
1	t	.139857	e	.122486	e	.259216
2	w	.089016	a	.115716	t	.117966
3	s	.088766	i	.096613	r	.093309
4	h	.079308	o	.089503	y	.090288
5	c	.060967	r	.074818	n	.072773
6	m	.060880	h	.065504	h	.071692
7	a	.055558	n	.057845	d	.063277
8	p	.052553	t	.055987	s	.045439
9	f	.050972	l	.051529	l	.042520
10	b	.047608	u	.048599	m	.038000
11	r	.037545	s	.038642	k	.026181
12	l	.034530	c	.038107	g	.020707
13	e	.032685	v	.030241	w	.014416
14	g	.029994	m	.023487	o	.012077
15	d	.027690	p	.016436	p	.010279
16	i	.020588	g	.016296	f	.006968
17	o	.018880	d	.014560	a	.006837
18	n	.016847	k	.009320	c	.004000
19	k	.013337	f	.008601	b	.002173
20	y	.011314	b	.008154	x	.000839
21	v	.009297	w	.007493	i	.000530
22	u	.008798	y	.004212	u	.000321
23	j	.008651	x	.003001	z	.000159
24	q	.004043	z	.001149	q	.000022
25	z	.000309	q	.000926	v	.000012
26	x	.000006	j	.000775	j	.000000

Note: letter counts were weighted with the Kucera & Francis (1967) frequency counts of the words they appeared in.

Words were selected for the four conditions to simultaneously satisfy two constraints. First, the means of the word frequencies in the high- and low-feature frequency conditions were matched. Second, the means of the feature frequencies A of the high- and low-frequency words were matched. In addition, each of the four conditions included approximately equal numbers of 4-, 5-, 6-, and 7-letter words. Since the range of feature frequency A is different for different word lengths, the matching was performed separately for the 4, 5, 6 and 7 letter words. We also verified that the words selected were still matched in feature frequency when we used feature frequency B as a measure. The means and standard deviations of the word frequencies, and feature frequencies A and B are listed for the four conditions in Appendix A2.

Each study list consisted of 130 words: 24 words from each of the four conditions and 34 filler items. Study position was randomly determined for each word for each subject, except for the first five words and the last five words, which were always filler items. Twelve targets and 12 distractors selected randomly from each condition were randomly assigned a serial position on the 96-item test lists.

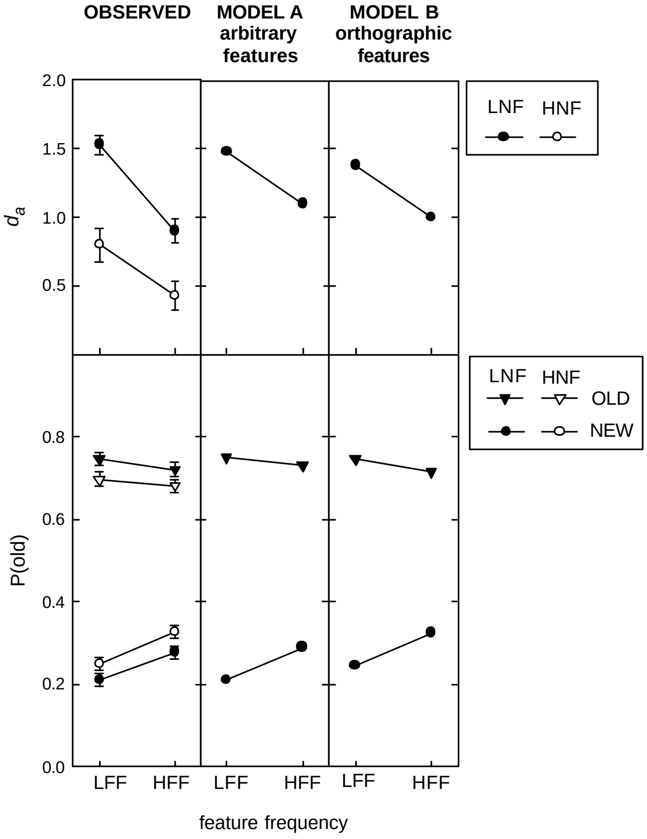


Figure 1. The results of the Experiment varying feature frequency and word frequency are shown in the left panels. The predicted results of model A and model B are shown in the middle and right panels respectively. The sensitivity results d_a are shown in the upper panels while the hit and false alarm rates are shown in the lower panels.

Procedure. An experimental session consisted of two study-test cycles. Participants were instructed prior to each study-test cycle to remember the words on the study list for a later memory test. Each word was displayed in uppercase form in the center of the computer screen for 1.3 s. of study. At test, participants performed a series of single-item ratings. Test items were presented one at a time, and participants were instructed to rate how confident they were that a test item was studied by utilizing a 6-point scale (a 1 indicated high confidence that an item had not been studied and a 6 indicated high confidence that an item had been studied). Responses were made by utilizing a mouse to click the appropriate button in the computer display. Each response was followed immediately by the presentation of a new item. At the end of the experiment, participants were given feedback concerning their performance on the task.

Results

The 6-point confidence ratings were converted to binary ‘old’-‘new’ responses by choosing a criterion and marking ratings higher or equal to the criterion as ‘old’ responses and ratings lower than the criterion as ‘new’ responses. For each participant, a criterion was chosen to equalize the overall number of ‘old’ and ‘new’ responses as much as possible[1]. The confidence ratings were used to compute ratings z-ROC curves by plotting the z transformed hit and false alarm rates using five criteria (1.5, 2.5, 3.5, 4.5 and 5.5) that were spaced between the confidence ratings. The z-ROC curves for each subject for each condition were used to compute sensitivity, d_a (Macmillan & Creelman, 1991; Swets & Pickett, 1982). An alpha of .05 was the standard of significance for all statistical analyses. In Figure 1 (left panel), d_a is shown for the four conditions in the top left panel. In the lower left panel, the mean probability of responding “old” is shown for the targets and distractors in the four conditions.

Word-frequency Effects. A typical word-frequency effect was observed. Mean d_a was greater for low-frequency than for high-frequency words [F(1,52) = 45.78, MSE = .42]. Hit rates were significantly higher for low-frequency words than for high-frequency words [F(1,52) = 11.77. MSE = .01], and the false-alarm rates were significantly lower for low-frequency words than for high-frequency words [F(1, 52) = 11.65, MSE = .01].

Feature-frequency Effects. Words consisting primarily of low-frequency letters were better recognized than words consisting primarily of high-frequency letters. Mean $\underline{d}_a$ for low feature-frequency words was greater than for high feature-frequency words [$\underline{F}$(1, 52) = 103.2, MSE = .13], and the interaction between word and feature frequency factors was significant [$\underline{F}$(1, 52)=4.47, MSE=.21]: the feature frequency effect was larger for low than high frequency words. Hit rates showed a small trend to be higher for words with low-frequency words than for words with high-frequency features [$\underline{F}$(1, 52) = 2.56, MSE = .01, p = 0.12], and the false-alarm rates were significantly lower for words with low-frequency features than words with high-frequency features [$\underline{F}$(1, 52) = 31.10, MSE = .01].

Discussion

The results confirm the prediction made by the REM model: words composed of primarily low frequency letters should be recognized better than words with primarily high frequency letters when the word frequencies are matched. The results also show that independent of feature frequency, at least as we measured this variable, word frequency also has a significant effect on performance: low frequency words are recognized better than high frequency words even if the feature frequencies of the words are matched. This suggests that feature frequency is one but not the only factor underlying the word frequency effect. Of course, feature frequency and other explanations for word frequency effects as mentioned in the Introduction are not mutually exclusive.

It is in principle possible that other word variables correlate with the feature frequency manipulation and that these other variables are causing the effects. Several variables such as concreteness and number of associations do not (wholly) explain the word frequency effect (Gorman, 1961; Kinsbourne & George, 1974), but could along with a potentially unlimited number of other variables (e.g. emotionality, imagery) correlate with the feature frequency manipulation. It would be no easy matter to explore such possibilities. An advantage of the present account is that feature frequency is easy to quantify objectively, and is easy to incorporate in a theoretical framework (as was done in REM).

Model Fits

REM uses Bayesian principles to model the decision process in recognition memory. This model as described by Shiffrin and Steyvers (1997, 1998) assumed events are represented as vectors of feature values, that episodic storage consists of forming incomplete and error prone copies of such events, that memory probes consist of vectors of feature values, and that retrieval is based on parallel matching of the features of the probes to the features of each memory trace. The matches and mismatches for each trace contribute evidence to a likelihood ratio for each trace and the odds for 'old' over 'new' turns out to be the sum of the likelihood ratios divided by the number of traces. This model was fit qualitatively to data from recognition memory experiments. Later, Diller, Nobel, and Shiffrin (in press) fit the model quantitatively to recognition and cued recall experiments. Even more recent work extended the model to various implicit memory tasks (e.g. Schooler, Shiffrin, & Raaijmakers, in press) and short-term priming (Huber, Shiffrin, Lyle, Ruijs, in press).

We modeled the results from this study in two ways. The first model, based on the REM model described by Shiffrin and Steyvers (1997), represents words with vectors of arbitrary feature values. The second model uses vectors of features based on the actual orthography of words, allowing the model to simulate performance by using models of the same words used in the experiment.

We opted not to model the effects of word frequency. Although the results showed effects of word frequency that were independent of feature frequency, there are many candidate mechanisms to model these additional effects, as described in the Introduction, and we are not ready to choose between these.

Model A, arbitrary features

In our first REM model, a vector of feature values, V, represents each word. The features are assumed to represent various attributes of words such as orthography (the number of features was set to 5). The values differ in their environmental base rates where the probability of choosing a feature value $\underline{V}$ is determined by the geometric distribution, based on a parameter, g:

$$P[V = j] = (1-g)^{j-1} g, \quad j = 1, \ldots, \infty \qquad (1)$$

The parameter $\underline{g}$ determines how common the average feature values drawn from the distribution will be: increasing $\underline{g}$ leads to word vectors with more common and less variable feature values.

To simulate the experiment for each subject, a lexicon of LFF and HFF words was generated to serve as target and distractor words. The stimulus vectors for the LFF and HFF conditions were generated with base rate parameters $\underline{g}_{LFF}$ and $\underline{g}_{HFF}$ respectively where $\underline{g}_{LFF} < \underline{g}_{HFF}$. Thus, LFF features are less common than the HFF features. To give an example, if we set $\underline{g}_{LFF}$=.1 and $\underline{g}_{HFF}$=.8, then (exaggerating a bit for the sake of the example) two likely word vectors for the LFF condition are

[9,4,14,25,6] and [7,27,2,15,8] and two likely word vectors for the HFF condition are [2,1,1,1,2] and [3,2,1,1,1]. Note that there are fewer features that overlap for the LFF vectors that the HFF vectors.

In REM, it is assumed that a separate image or trace is stored for each unique word studied. During study, copying feature values from the stimulus vectors to memory over occurs with a probability of c. With probability (1-c), a random feature is sampled from the geometric distribution defined by g_r and stored[2]. To simulate the experiment, 130 images were stored in memory, 65 LFF and 65 HFF words[3].

At test, a probe vector representing the test item is compared in parallel to all images in memory by counting the number of matching and mismatching features, m_j and q_j respectively, for each image, j,. For each probe-image comparison, a likelihood ratio λ_j is calculated:

$$l_j = (1-c)^{n_j} \prod_{k \in M_j} \frac{c + (1-c) f(V_{kj})}{f(V_{kj})} \tag{2}$$

This expresses the ratio of the probability that the image j matches the probe vector over the probability that the image does not match the probe vector.

In Equation (2), M_j is the set of matching features for image j and V_{kj} is the kth feature value in image j. The value $f(V)$ is the probability that feature value V was stored by chance. In this model, $f(V)$ was set to $(1-g_r)^{V-1} g_r$, the geometric distribution of Equation 1 using g_r as the base rate parameter.

The decision "old" or "new" is based on the odds φ that the probe is "old" over "new". A decision "old" and "new" is made when φ is bigger than 1 and smaller or equal than 1 respectively. In Shiffrin and Steyvers (1997), it was shown that this odds is equal to the sum of the likelihood ratio's λ_j divided by the number of images n:

$$j = \frac{P(\text{"}old\text{"})}{P(\text{"}new\text{"})} = \frac{1}{n} \sum_{j=1..n} l_j \tag{3}$$

This model uses four parameters and we tried a few sets of parameter values to model the observed results qualitatively (g_{LFF} = 0.3; g_{HFF} = 0.4; g_r = 0.32. c = 0.75). The top panel of the middle column in Figure 1 shows that sensitivity, d_a, is predicted to be greater for words comprised of low frequency features than for words comprised of high-frequency features[4]. The lower panel of the middle column of Figure 1 shows that a mirror effect for feature frequency is predicted: hits rates are lower for HFF words than LFF words and false alarms rates are higher for HFF words than LFF words.

The model predicts higher average false alarms rates for HFF than for LFF words because they have more features in common and because access to memory is assumed to be global. As a result a HFF word will tend to match the images of other words to a greater degree than do LFF words, which leads to higher likelihood ratio's and higher odds. A lower hit rate for HFF than LFF words is predicted because when features match, it is possible that they match by chance. Matching feature values will increase the likelihood ratios in Equation (1) to the degree that it is unlikely that the features match due to chance. Thus, even though HFF targets will lead to more matches than LFF targets, the matching values for HFF words contribute less to the likelihood ratios than the matching values for LFF words.

Model B: orthographic features

In Model A, the vectors for the LFF and HFF words differed in their environmental base rates of feature values but otherwise, these feature values were arbitrarily related to the stimulus features. In model B, we attempted to model more closely the stimulus structure of the experiment by choosing a representation for the words that is directly based on the orthography of the words. This enables us to make specific predictions based on the stimulus materials employed in this study.

The coding for the words in the experiment is directly based on the relative frequencies listed in Table 1. The most frequent letter is encoded with feature value "1", the second most frequent letter with feature values "2", and so on. For example, the vector [10,2,7,1] represents the word "bane", and the word "ajar" is encoded as [7,26,2,3]. Note that the initial letter "a" in "ajar" is encoded by value 7 and that the third letter "a" is encoded by value 2 because we distinguish between relative frequencies for different letter positions. This representation is a simple way to represent the orthographic structure of the stimulus materials and to capture the differences between the LFF and HFF words used in the experiment. Note that the LFF word "ajar" has a rare feature "j" while the word "bane" mostly consists of common features. The feature frequency differences in the stimulus materials will be reflected in the coding of the words, because common letters will be encoded by common feature values while rare letters will be encoded by rare feature values.

The same procedure for creating images was used as in model A. Error prone images of the study word vectors were created by storing the correct feature value with probability c. With probability (1-c), a random feature value was stored by sampling from the distribution of letter frequencies listed in Table 1. This is an empirical distribution of letter frequencies as they occur in the learning environment of an

English speaker. Because an explicit representation for words was available, the structure of the study list could be modeled: the 24 words from each of the four conditions and 24 filler items formed the 130-item study list.

At test, the probe vector was compared in parallel to each image in memory, and the number of matches and mismatches were calculated for each probe-image comparison. Because most of the probe and images consist of an unequal number of features, a choice has to be made of how to align the vectors and count matches and mismatches. A simple procedure was used in which the words were aligned at the beginning and ending of each word, and the best alignment in terms of number of matching features was chosen. Also, the difference in the number of features counted toward the number of mismatching features. For example, [1, 2, 3, 4, 5] and [6, 3, 4, 5] would have a best alignment at the end of the word and would give 3 matching features and 2 mismatching features (one due to the length mismatch). Other comparison procedures were also tried (such as no alignment at the end of the word or not counting the length mismatch between words) and gave qualitatively similar results.

With the number of matches and mismatches available, Equation (2) was applied to calculate the likelihood ratios for each image. The function $f(V)$ calculates the probability of matching the feature value V by chance, and its value is dependent on the relative feature frequencies listed in Table 1. Let $h(V)_p$ denote the relative frequency of letter V in position p of the word (first, interior or last). Then, we set $f(V)_p$ (it will be indexed with p because it will now also depend on letter position) not equal to $h(V)_p$ but on a less skewed distribution according to:

$$f(V)_p = \frac{\left[h(V)_p\right]^a}{\sum_{j=1..26}\left[h(V)_p\right]^a} \qquad (4)$$

where the parameter a determines the (un)skewing of the empirical distribution $h(V)_p$. We set $a<1$, to make the frequencies of the common and rare letters more similar. We will discuss this aspect of the model in more detail in a moment.

In Figure 1, left panels, the predicted results are shown for model B. With only two parameters, ($c=0.5$, $a=0.6$), this model can make predictions that are similar to both the observed data and the predicted data from model A. It predicts a mirror effect for the false alarm and hit rate for the same reasons as mentioned for model A: HFF words have more common features so HFF probes tend to match more features by chance which increases the false alarm rate. At the same time, there is a compensating factor that the common matching features will increase the likelihood ratio's less than rare matching features. The tradeoff between these two factors gives the predicted mirror effect.

This model can predict the effect of feature frequency based on a vector representation that is directly related to the stimulus material of the experiment and to the environmental base rates of the letters. Interestingly, to make this model work, it was necessary to make the environmental base rates less extreme so that the rare features were not as rare and common features were not as common[5]. One way to justify setting the base rates used by the model to values less extreme than the environmental base rates is based on the structure of the study list. Because the study list consists of many LFF words, the occurrence of rare letters such as “j”, “z” and “x” is less rare than outside this experimental setting. Participants might adjust their base rates to reflect these changes so that a “j”, “z” or “x” is less surprising than the environmental base rates suggest.

Another justification is based on work by Schooler and Anderson (1997) who argued and shown that rare items or features tend to clump together when they do occur: for example, a rare word seldom occurs, but when it occurs, it tends to reoccur shortly thereafter with a much higher probability than that given by the base rate. E.g., ‘flan’ seldom occurs but when it does it might do so because of a cooking context and would tend to reoccur. A generalization of this argument might be used to justify the higher than normal clumping of rare features generally (e.g. a scientific text might contain many rare feature values). If such clumping occurs, then the conditional probability that a rare feature value has been encountered recently, given that it is presented (in this case, for test) is much higher than the overall base rates would suggest.

Conclusion

Several recent global matching memory models explain the word frequency effect (Dennis & Humphreys, in review; Estes, 1993; Hintzman, 1997; McClelland & Chappell, 1998; Murdock, 1997) for a variety of reasons. This study suggests that these memory models need a component for feature frequency to explain part of the word frequency effect. In this article, we accounted for the feature frequency effects by assuming that the features that represent words differ in their base rates and that the recognition memory performance depends on these base rates: rare features are more diagnostic in the matching of the probe to the contents of memory than common features so performance is better for words with rare features than words with common features.

The first REM model assumed that the features of words comprised of primarily high or low frequent letters are represented by arbitrary features differing in their base rates. The second model employed a simple representation with which the letters of the experimental words were directly represented. Also, this model assumed that the diagnosticity of the features were directly dependent on the environmental letter frequencies.

There is one way in which the differences in feature frequency can be explained without using differences in the representation but rather differences in the amount of attention paid to words comprised of low and high frequent features. Participants might pay more attention to words with unusual features so that the encoding for the words with unusual features is better than words with common features. In this hypothesis, it still needs to be explained why participants pay more attention to words with unusual features in the first place. Second, implementing this idea in a model like REM by assuming that words with uncommon features lead to images with more features than words with common features leads to the prediction that the hit rates are affected by feature frequency but not the false alarm rates. In such a model, differences in false alarm rates can only be predicted if the participants can adjust the familiarity calculations (or an internal criterion) for probes (old or new) based on a guess as to what the encoding strength would have been were the probe stored in memory. Regardless of the plausibility of such assumptions, in order to model the experimental results based on differences in attention, a theory is needed in which feature frequency plays a central role because participants are assumed to notice differences in feature frequency and are assumed to adjust the familiarity calculations based on feature frequency.

Footnotes

Footnote 1. An alternative procedure is to use one criterion for all subjects such as the criterion between the first three and last three confidence ratings. With this alternative procedure all statistical results remain qualitatively the same. We choose the procedure of selecting criteria separately for each subject for two different reasons. First, this procedure correct for idiosyncratic use of the confidence scale (i.e., some participants use one end of the scale more than other participants). Second, a participant specific criterion leads to smaller standard errors in sensitivity, hits and false alarms than a universal criterion.

Footnote 2. In the Shiffrin and Steyvers (1997) REM model, there was an additional storage variable U*. This influenced the number of features that would be copied over from the probe and uncopied features were represented by the zero feature values. This variable was needed to explain study time and number of repetitions manipulations. Since this experiment did not involve these manipulations, we omitted this variable by assuming that all features of the words were stored.

Footnote 3. The experiment had 34 filler items and we choose not to model these separately and replaced them by 17 LFF and 17 HFF words.

Footnote 4. In order to compute d_a, five criteria were chosen (e^{-2},e^{-1},e^{0},e^{1},e^{2}) and hits and false alarms were computed to construct a z-ROC curve.

Footnote 5. Using the original base rates for f(V) or equivalently, setting the parameter a=1 in Equation (4) had the interesting effect that the false alarm rate for LFF words was higher than for HFF words. This is because a "new" LFF probe such as VORTEX contains the letter "x" and the letter "x" occurs in several other LFF images (e.g. PREFIX). The matching "x" contributes to a large increase in the likelihood ratio.

References

Brown, J. (1976). An analysis of recognition and recall and of problems in their comparison. In J. Brown (Ed.), Recall and Recognition. NY: Wiley.

Burnage, D. (1998). CELEX: a guide for users. Nijmegen, Netherlands: Centre for Lexical Information.

Chalmers, K.A., Humphreys, M.S., & Dennis, S. (1997). A naturalistic study of the word frequency effect in episodic recognition. Memory and Cognition, 25, 780-784.

Diller, D. E., Nobel, P. A., and Shiffrin, R. M. (in press). An ARC-REM
model for accuracy and response time in recognition and recall. Journal of Experimental Psychology: Learning, Memory, and Cognition.

Glanzer, M., & Adams, J.K. (1990). The mirror effect in recognition memory: data and theory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 16, 5-16.

Glanzer, M., & Bowles, N. (1976). Analysis of the word-frequency effect in recognition memory. Journal of Experimental Psychology: Human Learning and Memory, 2 (1), 21-31.

Gorman, A. N. (1961). Recognition memory for names as a function of abstractness and frequency. Journal of Experimental Psychology, 61, 23-29.

Gregg, V.H. (1976). Word frequency, recognition, and recall. In J. Brown (Ed.), Recall and recognition. London: Wiley.

Huber, D. E., Shiffrin, R. M., Lyle, K. B., & Ruys, K. I. (in press). Perception and preference in short-term word priming. Psychological Review.

Kinsbourne, M., & George, J. (1974). The mechanism of the word frequency effect on recognition memory. Journal of Verbal Learning and Verbal Behavior, 13, 63-69.

Kucera, H., & Francis, W.N. (1967). Computational analysis of present-day American English. Providence, RI: Brown University Press.

Landauer, T.K., & Streeter, L.A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. Journal of Verbal Learning and Verbal Behavior, 12, 119-131.

Lockhart, R.S., Craik, F.I.M., & Jacoby, L. (1976). Depth of processing, recognition, and recall. In J. Brown (Ed.), Recall and Recognition. NY: Wiley.

Macmillan, Neil A. and C. Douglas Creelman. (1991). Detection Theory: A user's guide. Cambridge, England: Cambridge University Press.

McClelland, J.L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. Psychological Review, 105, 724-760.

McCormack, P. D., & Swenson, A. L. (1972). Recognition memory for common and rare words. Journal of Experimental Psychology, 95, 72-77.

Scarborough, D., Cortese, C., & Scarborough, H. (1977). Frequency and repetition effects in lexical memory. Journal of Experimental Psychology: Human Perception and Performance, 3, 1-17.

Schooler, L. J., & Anderson, J. R. (1997). The role of process in the Rational Analysis of Memory, Cognitive Psychology, 32, 219-250.

Schooler, L .J., Shiffrin, R. M., & Raaijmakers, J. G. W. (in press). Theoretical note A Bayesian model for implicit effects in perceptual identification. Psychological Review.

Shepard, R.N. (1967). Recognition memory for words, sentences, and pictures. Journal of Verbal Learning and Verbal Behavior, 6, 156-163.

Shiffrin, R.M., & Steyvers, M. (1997). A model for recognition memory: REM – retrieving effectively from memory. Psychonomic Bulletin & Review, 4, 145-166.

Shiffrin, R. M., & Steyvers, M. (1998). The effectiveness of retrieval from memory. In M. Oaksford & N. Chater (Eds.). *Rational models of cognition*. (pp. 73-95), Oxford, England: Oxford University Press.

Schulman, A. I., & Lovelace, E. A. (1970). Recognition memory for words presented at slow or rapid rate. Psychonomic Science, 21, 99-100.

Zechmeister, E. B. (1969). Orthographic distinctiveness. Journal of Verbal Learning and Verbal Behavior, 8, 754-761.

Zechmeister, E. B. (1972). Orthographic distinctiveness as a variable in word recognition. American Journal of Psychology, 85, 425-430.

Appendix A
Words of Experiment 1

LFF-LNF

ABLAZE
ACRYLIC
AJAR
ALFALFA
APEX
AVOCADO
AVOW
AZALEA
BOXING
CHIMP
CHOMP
CHUBBY
CONVEX
DYNAMIC
ELYSIUM
ENCAMP
EPIC
EPOCH
ERGO
EXCERPT
EXHALE
EXHAUST
FLUX
GAWKY
GUSTO
HUMP
IMPEL
JAGGED
JOGGING
JOWL
JUNO
KILO
KIOSK
KNACK
KNOBBLY
KNOWING
LIEU
LOCKS
LYRICS
MAYFLY
MIDRIFF
NOVA
NUMBLY
ODYSSEY
OOZE
OPOSSUM
OUTBACK
OUTGROW
OZONE
PREFIX
PSYCHE
PUFFY
QUAKE
QUIBBLE
QUICKEN
QUIP
QUIRK
REVAMP
SKIMP
SQUID
STANZA
SWAB
TWITCH
TYPHOID
UPTIGHT
UTOPIA
VERB
VIVA
VORTEX
WHACK
YANK
YOLK

HFF-LNF

ALERT
BANE
BARTER
BASTE
BEET
BROILER
BRUTE
CALLER
CENSURE
COERCE
CURLY
CURRANT
DALE
DEAREST
DECREE
FAINT
FERRET
FLIER
GALORE
LEARNER
PARROT
PASTE
PATE
PATRIOT
PEAT
PETITE
PLIANT
PORE
RELIANT
RILE
SEARING
SEDATE
SENSORY
SHEAR
SHINE
SOLID
SOOT
SPORE
STEROID
STRUT

BILE
BOILER
BRAID
BRAY
COOLER
CORNET
CORONER
COTE
DELETE
DILATE
DINER
DIRE
MANE
MARINER
MIRE
PALETTE
PELLET
PENAL
PENANCE
PERT
SAIL
SAUCY
SAUNTER
SCARLET
SILT
SINNER
SMEAR
SNOOTY
SUNRISE
TANNERY
TENSE
TINE

LFF-HNF

AMAZING
ATOMIC
AWFULLY
AWKWARD
BUREAU
CLIFF
CLIMB
COMPLEX
DIFFER
DOZEN
EGYPT
ELBOW
EVOLVE
EXAM
EXCEED
EXCLAIM
EXERT
EXIT
EXPLODE
EYEBROW
GHETTO
GOLF
GULF
HAZARD
INDEX
INJURY
JACKAL
KICK
KINGDOM
KNIGHT
LAMB
LIMB
LIQUID
LOBBY
LOGIC
LUXURY
MAJOR
MIXED
MYTH
NATO
NETWORK
ODDS
OFFEND
OMEGA
OPERA
OTTO
OXYGEN
PHOTO
PHYSICS
PUZZLED
RHYTHM
RUBBER
SYMBOL
SYMPTOM
TAXI
THIGH
THOU
THUMB
TOBACCO
TOMB
UNDERGO
UNHAPPY
UNIFORM
UNIQUE
UNKNOWN
UPWARDS
VACUUM
WAYS
WHIP
WHISKY
WIDOW
ZERO

HFF-HNF

AIRLINE
BAIT
BALLET
BARREL
BARRIER
BEAR
BEAST
BETRAY
BITE
BLEED
CANAL
CATTLE
CELLAR
CLAY
CLIENT
CORE
CORRECT
CRUELTY
CURE
CURRENT
DAISY
DEALER
DENSE
FARE
FLEET
FREE
GALLERY
GREET
MALE
MINER
MINERAL
MIRACLE
PAINTER
PANEL
PARADE
PEASANT
PENALTY
PILE
PILOT
PINE
PLAIN
PLANET
PLANNER
PLEAD
POET
POLE
PRAY
PRESENT
RALLY
RELATE
RELEASE
RETIRE
SAME
SCENT
SEAL
SECURE
SENATOR
SHEER
SHORE
SPINE
STARTLE
STATUE
STORAGE
STRAIN
TALE
TENURE
TERRACE
TERROR
TOILET
TRACE
TRAY
TREATY

Note. LFF-LNF = low orthographic feature frequency, low normative word frequency; HFF-LNF = high orthographic feature frequency, low normative word frequency; LFF-HNF = low orthographic feature frequency, high normative word frequency; HFF-HNF = high orthographic feature frequency, high normative word frequency.

Appendix B
Means and standard deviations of the word frequencies and feature frequencies A and B

		FF Condition			
Measure	WF Condition	LFF		HFF	
Word Frequency	LWF	4.10	(1.22)	4.56	(1.41)
	HWF	23.56	(6.61)	25.29	(7.17)
Feature Frequency A	LWF	2.17E-7	(3.38E-7)	2.18E-5	(3.59E-5)
	HWF	2.17E-7	(4.21E-7)	2.21E-5	(3.66E-5)
Feature Frequency B	LWF	.0501	(.0173)	.0770	(.0163)
	HWF	.0531	(.0167)	.0805	(.0194)

Note: standard deviations are given between parentheses