Memory & Cognition
1987, 15(1), 24-28

The word frequency effect in lexical decision: Finding a frequency-based component

MICHAEL K. GARDNER
University of Utah, Salt Lake City, Utah

E. Z. ROTHKOPF
AT&T Bell Laboratories, Murray Hill, New Jersey

and

RICHARD LAPAN and TOBY LAFFERTY
University of Utah, Salt Lake City, Utah

Subjects making lexical decisions are reliably faster in responding to high-frequency words than to low-frequency words. This is known as the word frequency effect. We wished to demonstrate that some portion of this effect was due to frequency differences between words rather than to other dimensions correlated with word frequency. Three groups of subjects (10 engineers, 10 nurses, and 10 law students) made lexical decisions about 720 items, half words and half nonwords, from six different categories (engineering, medical, low-frequency nontechnical, medium-frequency nontechnical, and two groups of high-frequency nontechnical). Results of two analyses of variance revealed a crossover interaction such that engineers were faster in responding to engineering words than to medical words, whereas nurses were faster in responding to medical words than to engineering words. The engineering and medical words were equally long and equally infrequent by standard word counts. We take this as support for a frequency-based component in the word frequency effect. The practical implications of this research for estimating the readability of technical text are discussed.

The lexical decision task is commonly used to study the characteristics of the internal lexicon (McKoon & Ratcliff, 1979; Meyer & Schvaneveldt, 1971; Meyer, Schvaneveldt, & Ruddy, 1974; Rubenstein, Garfield, & Millikan, 1970; H. Rubenstein, Lewis, & M. Rubenstein, 1971; D. L. Scarborough, Cortese, & H. S. Scarborough, 1977; Schvaneveldt, Meyer, & Becker, 1976). In one version of such a task, a subject must decide whether a string of visually presented letters is a word or a nonword. Subjects making this decision are reliably faster in responding to high-frequency words than they are in responding to low-frequency words (Rubenstein et al., 1970; Scarborough et al., 1977). This finding is known as the word frequency effect.

It would seem that the difference in time required to respond to high- and low-frequency words is due to the differing frequency of these items in the general language. This would imply that people making a lexical decision are influenced by their experience with the lexicon, and that greater experience leads to faster decisions. Upon reflection, however, we found that another possible explanation of the word frequency effect exists. Landauer and Streeter (1973) showed that a number of lexical dimensions are correlated with word frequency. For instance, common and rare words have different distributions of phonemes and graphemes. A number of other differences exist as well. Landauer and Streeter pointed out that any of these differences could account for the word frequency effect. They stated:

> The observed differences in phonemic constituency, associated with apparent differences in communicative effectiveness, suggests [sic] a route by which characteristics other than frequency may account for the frequency effect. But the size of these effects is very difficult to evaluate relative to the size of the word-frequency effect itself. It would be hard to show that all or any specifiable portion of the word-frequency effects usually observed are due to such factors. (p. 130)

We wished to demonstrate that at least some portion of the word frequency effect is due to frequency—that is, experience with the lexical items in question—and not to some confounding factor, such as those suggested by Landauer and Streeter. To test our contention, we performed the experiment described here. The rationale was to ask

This research was supported jointly by AT&T Bell Laboratories and a grant to the first author from the Graduate School of Education of the University of Utah. The first experiment was conducted while the first author was a summer research fellow at AT&T Bell Laboratories, Murray Hill, NJ. The second author is now at both AT&T Bell Laboratories and Columbia University Teachers College. The authors wish to thank Dennis Egan, Peter Dixon, and Jerry Billington of AT&T Bell Laboratories, as well as John Cartan of the University of Utah, for their help with various aspects of this study.

Requests for reprints can be addressed to: Michael K. Gardner, Department of Educational Psychology, 327 Milton Bennion Hall, University of Utah, Salt Lake City, UT 84112.

members of two different occupations to make lexical decisions about two groups of equally infrequent, occupationally related words. We hoped to find an interaction between the occupations of subjects and the categories of words, such that subjects were relatively fast at responding to words from their own occupation, but relatively slow at responding to (equally infrequent) words from another occupation. Such results could be explained in terms of the greater frequency with which members of an occupation encounter technical words from their own occupation, even though, on the basis of standardized word counts (e.g., Kučera & Francis, 1967), such technical words are of quite low frequency. This would support the notion that at least some portion of the word frequency effect is due to frequency—that is, the number of times an individual encounters a word in his or her day-to-day experience.

METHOD

Subjects

Subjects were selected from three groups with differing technical backgrounds. All three groups of subjects came from the Salt Lake City, Utah, metropolitan area. The first group consisted of 10 graduate students and professionals (9 males and 1 female) with backgrounds in engineering, physics, or computer science. The second group consisted of 10 graduate students and professionals (1 male and 9 females) with backgrounds in nursing. The third group, which served as a university-level control group, consisted of 10 law students (7 males and 3 females). The students in each of the three groups were enrolled in graduate programs at the University of Utah. The professionals in engineering, physics, and computer science had been trained in these areas and were now employed in their respective fields. The professionals in nursing had been trained in nursing and were employed at the time of the experiment as registered nurses. All subjects were paid for their participation in the study.

Stimuli

Stimuli in the experiment consisted of six groups of 120 words. All of the words in each of the six groups were nouns or were used primarily as nouns.

The first group was taken from the indices of a number of texts on electrical engineering and physics. The second group was compiled from the index of a textbook on medical and surgical nursing. These items constituted the two groups of technical words in the experiment: engineering words and medical words.

The third and fourth groups were selected from high-frequency words found in the Kučera and Francis (1967) word-count norms. High frequency was defined as a frequency of 100 per million or greater. The fifth and sixth groups consisted of medium- and low-frequency words from the Kučera and Francis norms. Medium frequency was defined as a Kučera-Francis frequency count of between 10 and 99 occurrences per million items of text, and low frequency was defined as a Kučera-Francis count of fewer than 10 occurrences per million items of text. Words in groups three through six did not represent any specialized occupational background. These groups served two purposes: (1) they allowed lexical decisions about letter strings to be made against a background of items of varying frequencies, and (2) they allowed a number of baselines against which to measure the size of the word frequency effect.

Half of the words in each set were converted to nonwords by changing one or two letters. Vowels replaced vowels and consonants replaced consonants. All resulting nonwords were pronounceable and followed the rules of English orthography.

Median-frequency counts (using the Kučera-Francis norms), as well as mean string lengths, for each of the six sets of stimuli (broken down by words, and the words on which the nonwords were based) are presented in Table 1. Median frequency counts, rather than mean frequency counts, are reported because the distribution of Kučera-Francis word frequencies is skewed in the three low-frequency groups (engineering, medical, and low-frequency nontechnical). This skewing occurs because the distribution is bounded on the low end by one (the lowest possible frequency per million reported by Kučera and Francis). We also calculated the mean frequency count for each of the stimulus groups listed in Table 1, and they displayed a pattern very similar to the pattern displayed by the medians. As can be seen from the table, words in both of the technical word groups are very infrequent and of roughly equivalent string length. Words in the high-frequency groups are very frequent, and of somewhat shorter string length. Words in the low-frequency nontechnical group are similar to those of the two technical groups in terms of median frequency and mean string length, and words in the medium-frequency group are intermediate between those of the three low-frequency groups and the two high-frequency groups.

It is worth noting that there is a substantial negative correlation between string length and word frequency across the six stimulus groups in our sample. This correlation between string length and frequency occurs in the language, and is very difficult to avoid without producing biased samples of words. For this reason no attempt was made to equate the six stimulus groups in terms of string length. Although this correlation makes it difficult to eliminate string length as a potential contributor to differences in lexical decision times among some of the groups, the two groups of primary interest—the technical groups—are matched in terms of string length. Thus differences in lexical decision time between the two technical groups should not be affected by differences in string length.

In addition to the six groups of items discussed above, a group of 60 items (30 words and 30 nonwords) was constructed for the purpose of practice. These items were chosen so as not to represent any particular frequency group or technical background; they merely served to familiarize the subjects with the experimental task. Lexical decision times to these items were not included in any of the analyses to be discussed.

Design

Subjects made a word-nonword decision on all 720 stimulus items (plus the additional 60 practice items). The 60 practice items were blocked together and were always the first set of items presented.

Table 1
Frequency Counts and String Lengths of Experimental Stimuli

Category	Lexical Status	Kučera-Francis Frequency Count Median	String Length Mean	String Length *SD*
Engineering	Word	2.5	8.30	2.29
	Nonword	3.5	8.42	2.31
Medical	Word	1.0	8.25	2.24
	Nonword	1.0	8.30	2.15
High-Frequency Group 1	Word	145.0	6.72	2.37
	Nonword	185.5	6.63	2.31
High-Frequency Group 2	Word	159.5	6.80	1.38
	Nonword	158.5	6.85	1.64
Medium Frequency	Word	36.5	7.27	1.54
	Nonword	27.0	6.80	1.90
Low Frequency	Word	2.0	8.58	2.06
	Nonword	2.0	7.73	2.47

Note—Statistics for nonwords are based on the words from which they were created.

The 720 stimulus items were presented in 12 blocks of 60 items each. Each block contained 5 words and 5 nonwords from each of the six stimulus categories (engineering, medical, high-frequency Group 1, high-frequency Group 2, medium frequency, and low-frequency nontechnical). The 60 items in each block were chosen randomly without replacement from the total pool of items. For instance, in forming Block 2 the 5 high-frequency words could be any of the high-frequency words except those 5 that had already served in Block 1. Arrangement of items within a block was also random, with the constraint that no more than 5 words or 5 nonwords be presented consecutively.

Procedure

All stimulus presentations and data collection were controlled via an Apple II+ microcomputer. Subjects were seated in a room illuminated by subdued light, in front of a 12-in. black and white television monitor. Subjects responded by pressing either the z key (far left) or the ? key (far right) on the keyboard of the computer. Subjects responded with their preferred hand to stimuli which they believed to be words and with their nonpreferred hand to stimuli which they believed to be nonwords.

At the beginning of a trial, an asterisk, centered in the presentation field, was presented for 1,000 msec as a fixation point and as a warning that a trial was about to begin. The screen then went blank for 1,000 msec, at the end of which time the stimulus string appeared in white capital letters on the darkened background of the television screen. When the subject made his or her response, the stimulus was erased and the screen remained blank for 400 msec. The subject's reaction time and the correctness of the response were recorded. Feedback was then provided for 500 msec; either the word *correct* or the word *incorrect* was displayed centered on the television screen. The screen was then cleared and remained blank for 500 msec preceding the beginning of the next trial.

At the conclusion of each block, the subject was given feedback on his or her mean reaction time (for correct trials only) and the number of errors he or she made. If the subject made more than eight errors, the following message was displayed: "Your error rate is high. Remember—accuracy is important." Subjects were instructed at the beginning of the session to respond as quickly as possible while maintaining high accuracy.

RESULTS

In this experiment we were interested in assessing the difficulty of a lexical decision when such a decision could in fact be made. Thus we will be reporting data for correct responses only. We performed similar analyses on the entire data set (correct plus incorrect decision times combined), and these analyses produced similar results.

Our experiment yielded five potentially relevant factors: (1) occupational background (engineer, nurse, or law student), (2) category of the stimulus item (engineering, medical, high-frequency Group 1, high-frequency Group 2, medium-frequency, and low-frequency nontechnical), (3) lexical status of the stimulus item (word or nonword), (4) presentation block (one through twelve), and (5) subject (10 within each profession). We began by collapsing data over the five replications that existed within the crossing of each level of each of the five factors. This yielded a more stable estimate for each cell. We next transformed our data to reduce the effect of outliers on later analyses. For each subject, we calculated the standard deviation of lexical decision times. We then set any lexical decision time longer than two standard deviations above the mean equal to two standard deviations above the mean. Finally, we collapsed our data over blocks, because this factor was of no interest to us.

The main results from the experiment are presented in Table 2, which shows lexical decision times for correct

Table 2
Lexical Decision Times for Correct Responses and Error Rates

Lexical Status		Category					
		Engineering	Medical	High-Frequency Group 1	High-Frequency Group 2	Medium Frequency	Low Frequency
				Engineers			
Word	Mean	790	946	679	667	728	907
	SD	188	300	136	141	158	276
	Error Rate	.07	.13	.01	.01	.02	.12
Nonword	Mean	1009	1065	918	929	932	978
	SD	367	399	289	320	303	353
	Error Rate	.05	.03	.02	.03	.05	.02
				Nurses			
Word	Mean	1025	926	677	692	757	1054
	SD	498	474	270	293	313	520
	Error Rate	.25	.02	.01	.01	.02	.12
Nonword	Mean	1125	1167	1007	1029	1000	1045
	SD	540	580	462	487	425	468
	Error Rate	.06	.04	.02	.05	.07	.04
				Law Students			
Word	Mean	764	756	554	552	593	732
	SD	152	114	98	91	93	139
	Error Rate	.20	.08	.01	.01	.00	.06
Nonword	Mean	854	908	774	778	802	813
	SD	176	235	127	143	146	150
	Error Rate	.05	.03	.02	.03	.03	.02

responses, the standard deviations of these lexical decision times, and error rates broken down according to the category and the lexical status of the stimulus. The results are further subdivided by the occupational backgrounds of the subjects.

Error rates for engineers and nurses were high for infrequent words, including occupational words from outside their own occupations (for engineers, low frequency = 12%, medical = 13%; for nurses, low frequency = 12%, engineering = 25%). For law students, who served as a control group, error rates were also high for the engineering words (20%), and somewhat high for the medical words (8%) and the low-frequency words (6%). Although the 8% and 6% error rates seem small in absolute terms, the law student group was extremely able verbally, and showed better overall performance than either the engineering or the nursing group. The results from the law students indicate that the engineering words were somewhat more difficult than the medical or low-frequency words, but that these two groups were also more difficult than any of the other three groups of words.

A four-factor analysis of variance was performed on the reaction time data. The four factors in the analysis were: (1) occupational background of subject, (2) category of the stimulus item, (3) lexical status of the stimulus item, and (4) subject. The first three of these factors were crossed; the first was a between-subjects factor, whereas the remaining two were within-subjects factors (repeated measures). These three factors were treated as fixed effects. The last factor was nested within the first and was treated as a random effect.

The results of the analysis of variance are reported in Table 3. The finding of primary interest is the three-way interaction of lexical status of the stimulus item × category of the stimulus item × occupational background of the subject [$F(10,135) = 4.36$, $p < .001$]. The significance of this interaction, we contend, is due in large part to the fact that nurses were (relatively) fast on medical words and slow on engineering words, whereas engineers were (relatively) fast on engineering words and slow on medical words.

The problem with interpreting the interaction of lexical status × category × occupation as we have done is that there are a number of cells that might be producing the interaction, but which are not part of our hypothesis. To clarify the situation, and to test our hypothesis more explicitly, we conducted a second analysis of variance. In this analysis, we restricted ourselves to considering only: (1) the categories of engineering and medical items, (2) the occupations of engineer and nurse, and (3) the lexical status of word (no effect was predicted for nonword decisions). The experimental hypothesis was tested by the interaction of category × occupation (because there is only one lexical status possible, there is no additional factor for lexical status). This interaction was significant [$F(1,18) = 27.25$, $p < .001$], which validates our earlier interpretation. Significant individual differences in lexical decision can be found, even when the word groups

Table 3
Analysis of Variance of Correct Lexical Decision Times

Variance Source	*df*	Sum of Squares	Mean Square	*F*
Occupation of Subject	2	2,936,413	1,468,206	1.30
Subjects within Occupation	27	30,384,964	1,125,369	
Lexical Status of Stimulus Item	1	3,086,914	3,086,914	75.36*
Lexical Status × Occupation	2	28,980	14,490	0.35
Lexical Status × Subjects within Occupation	27	1,105,964	40,962	
Category of Stimulus Item	5	2,262,481	452,496	50.95*
Category × Occupation	10	186,045	18,604	2.09†
Category × Subjects within Occupation	135	1,198,949	8,881	
Lexical Status × Category	5	552,451	110,490	30.60*
Lexical Status × Category × Occupation	10	157,600	15,760	4.36*
Lexical Status × Category × Subjects within Occupation	135	487,521	3,611	

*$p < .001$. †$p < .05$.

in question have very similar frequencies in the general language and nearly identical string lengths. What caused these individual differences? We conclude that the differences were due to the differing amounts of experience that each subject group had with each word group. This brings us back to the notion of word frequency. Although the two word groups had similar frequencies for subjects drawn randomly from the general populace, each word group was of higher frequency for subjects drawn from a technical background, who used these words as part of their occupational vocabulary. Thus we found a frequency effect based on experience: greater experience with a group of words leads to faster lexical decisions.

How large is the word frequency effect in this experiment? It depends, to some extent, upon the baseline against which we choose to measure it. If we use the two high-frequency groups as the baseline, then the word frequency effect, as measured by technical words outside of a subject's occupation, is 273 msec for engineers and 341 msec for nurses. If we use the same baseline, but this time measure the word frequency effect using technical words from within a subject's occupation, the corresponding numbers are 117 msec for engineers and 242 msec for nurses. This amounts to a 57% reduction in the word frequency effect for engineers and a 29% reduction for nurses. The size of the frequency-based component of the word frequency effect is substantial.

If the word frequency effect is measured from some other baseline, the size of the absolute word frequency effect (technical words from outside a subject's occupation minus the baseline) diminishes, but the proportional reduction due to a word's inclusion in a subject's occupational vocabulary increases. For instance, if the medium-frequency group is used as the baseline, the absolute word frequency effects are 218 msec for engineers and 268 msec for nurses. Measured from technical words within a subject's occupation, the word frequency effect is 62 msec for engineers and 169 msec for nurses. The reduction due to a word's inclusion in a subject's occupa-

tional vocabulary is now 72% for engineers and 37% for nurses. Thus our earlier calculations set a lower bound on the size of the frequency-based component of the word frequency effect.

DISCUSSION

The findings of the experiment are clear. Systematic individual differences in lexical decision times were found as a function of subjects' occupational backgrounds and the technical categories from which words were drawn. In particular, nurses were relatively faster in responding to medical words than to engineering words, whereas engineers were relatively faster in responding to engineering words than to medical words. The two groups of technical words, however, had nearly identical frequency counts in the Kučera and Francis (1967) word-count norms.

Our results demonstrate that a substantial portion of the word frequency effect is due to frequency—that is, experience with a corpus of words—and not to the many factors correlated with frequency, such as those demonstrated by Landauer and Streeter (1973). Although we acknowledge that such correlated factors exist and contribute to the word frequency effect, this experiment controlled for them by taking the same two groups of technical words and varying only the subject populations that viewed them. The result—substantial differences in lexical decision time—must be due to differences among the populations. The most obvious of these is a difference in educational background and work experience. This, we contend, is the heart of a true frequency-based difference. Word frequency, in an approximate way, reflects the familiarity of the subject with a word's meaning and the contexts in which it is likely to occur, and may also indicate how recently it has been seen.

This experiment also has practical implications for estimating the readability of technical expository text. Computerized automatic methods have been developed (e.g., Coke & Rothkopf, 1970; Macdonald, Frase, Gingrich, & Keenan, 1982) for calculating readability indices on the basis of a number of structured characteristics of text. Word frequency is one of these structured characteristics, although it is usually indirectly estimated through correlated measures such as word length (e.g., Flesch, 1948) or by use of special indicator lists of high-frequency words (Dale & Chall, 1948). As Miller and Kintsch (1980) have pointed out, readability can be viewed as an interaction between reader and text. The established approximations used in readability formulae are based on general word frequency norms. For this reason they may not be accurate estimates of the familiarity of specialized target populations with the words of the text, and may lead to underestimates of readability.

How can this problem be avoided? A readability predictor that is tailor-made for a particular group of readers could take several forms. It might, for example, make use of a discounting coefficient for the length of technical words, or calculate the lexical components of the readability formula directly from word counts in particular literatures. Whatever form such an amended readability predictor may take, it will have the interesting property of being less restrictive for writers and editors than conventional difficulty indices. For this reason, a readability index that makes use of expectations about the language experience of intended readers is likely to make the development of suitable written documentation less costly.

REFERENCES

Coke, E. U., & Rothkopf, E. Z. (1970). Note on a simple algorithm for a computer-produced reading ease score. *Journal of Applied Psychology*, **54**, 208-210.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability. *Educational Research Bulletin*, **27**, 11-20.

Flesch, R. F. (1948). A new readability yardstick. *Journal of Applied Psychology*, **32**, 221-233.

Kučera, H., & Francis, W. (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 & Verbal Behavior*, **12**, 119-131.

Macdonald, N. H., Frase, L. T., Gingrich, P. S., & Keenan, S. A. (1982). The Writer's Workbench: Computer aids for text analysis. *IEEE Transactions on Communications*, **Com-30**, 105-110.

McKoon, G., & Ratcliff, R. (1979). Priming in episodic and semantic memory. *Journal of Verbal Learning & Verbal Behavior*, **18**, 463-480.

Meyer, D., & Schvaneveldt, R. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, **90**, 227-234.

Meyer, D., Schvaneveldt, R., & Ruddy, M. (1974). Functions of graphemic and phonemic codes in visual word-recognition. *Memory & Cognition*, **2**, 309-321.

Miller, J. R., & Kintsch, W. (1980). Readability and recall of short prose passages: A theoretical analysis. *Journal of Experimental Psychology: Human Learning & Memory*, **6**, 335-354.

Rubenstein, H., Garfield, L., & Millikan, J. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning & Verbal Behavior*, **9**, 487-494.

Rubenstein, H., Lewis, S., & Rubenstein, M. (1971). Evidence for phonemic recoding in visual word recognition. *Journal of Verbal Learning & Verbal Behavior*, **10**, 645-657.

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

Schvaneveldt, R. W., Meyer, D. E., & Becker, C. A. (1976). Lexical ambiguity, semantic context, and visual word recognition. *Journal of Experimental Psychology: Human Perception & Performance*, **2**, 243-256.

(Manuscript received November 11, 1985;
revision accepted for publication July 7, 1986.)