WADC TECHNICAL REPORT 54-282

ON THE INTERPRETATION OF WORD FREQUENCY AS A VARIABLE AFFECTING SPEED OF RECOGNITION

DAVIS H. HOWES

AERO MEDICAL LABORATORY

JUNE 1954

WRIGHT AIR DEVELOPMENT CENTER

WADC TECHNICAL REPORT 54-282

ON THE INTERPRETATION OF WORD FREQUENCY AS A VARIABLE AFFECTING SPEED OF RECOGNITION

Davis H. Howes

Aero Medical Laboratory

June 1954

RDO No. 694-37

Wright Air Development Center
Air Research and Development Command
United States Air Force
Wright-Patterson Air Force Base, Ohio

Carpenter Litho & Prtg. Co., Springfield, O.
100 - 28 September 1954

FOREWORD

This report was prepared by the Psychology Branch of the Aero Medical Laboratory, Directorate of Research, Wright Air Development Center, under a project identified by Research and Development Order No. 694-37, "A Mathematical Approach to Human Communications," with Davis H. Howes acting as Project Engineer.

Analysis of the data was performed by Antioch College on Contract No. W33(038)-ac-19816.

ABSTRACT

Interpretations of the inverse relation between the duration for which a printed word must be exposed visually in order to be recognized and the frequency of occurrence of the word in a large general word count are examined critically. It is found that the most satisfactory interpretation is to regard the word's frequency in a word count as an estimate of its average probability of emission by the population of subjects used in the recognition experiment (base probability). Since the threshold of recognition is defined by the probability of emission of the word following its exposure, only a short exposure will be necessary to bring up to threshold criterion a word whose base probability is almost as great as the criterion probability, while a much longer exposure will be required for a word whose base probability is low. The critical assumption of this interpretation is that the frequency of a word in a large general word-count represents its average probability of emission by the experimental subjects used in the recognition experiment. Three experiments evaluating the validity of this assumption for the Lorge Magazine Count are described.

PUBLICATION REVIEW

This report has been reviewed and is approved.

FOR THE COMMANDER:

JACK BOLLERUD
Colonel, USAF (MC)
Chief, Aero Medical Laboratory
Directorate of Research

TABLE OF CONTENTS

	Page
Introduction	1
Procedure	2
Experiments	3
Experiment 1	3
Experiment 2	4
Experiment 3	5
Discussion	9
Summary	11
References	12

ON THE INTERPRETATION OF WORD FREQUENCY AS A VARIABLE AFFECTING SPEED OF RECOGNITION

INTRODUCTION

The duration for which a printed English word must be presented visually to a subject in order for him to recognize it is inversely correlated with the frequency of occurrence of the word in large samples of written English (2, 3, 4, 7, 8, 11). Since the former quantity (the duration threshold) is generally regarded as a perceptual variable and the latter (word frequency) as a response variable, this correlation offers a point of departure for the formulation of perceptual phenomena in behavioral concepts. The object of the present study is to test experimentally an assumption basic to one interpretation of this correlation. A mathematical formulation of the experimental data based upon this interpretation will be presented in a subsequent report.

The interpretation to be considered here can be characterized as a response-emission theory. We may think of the momentary probability of a word (defined as the strength of the subject's tendency to emit that word in preference to any other) as a quantity that fluctuates widely from moment to moment in accordance with changes in innumerable environmental and organismic conditions that affect the emission of words. Over a time period of considerable length the average of these momentary probabilities will be a relatively stable statistic, which we shall call the base probability of the word.

Visual exposure of a word to a subject for a brief length of time Δt is assumed to represent an environmental event tending to cause emission of the exposed word. The momentary probability of a word following its exposure may therefore be analyzed into two components: a component due to the ordinary impulses to emission of the word, whose average value is the base probability; and a component due to the additional impulse of the word's visual exposure. Consequently, the average probability of a word following each of a number of exposures of the same duration must be greater than the corresponding average base probability of the word. A given level of probability following exposure can result either from a relatively large component due to base probability plus a small additional component due to exposure or from a relatively small component due to base probability plus a large additional component due to exposure. It follows that the duration threshold of a word, which is defined as the duration of exposure for which 50% of the subject's reports following exposure are correct, will be lower for a word with high base probability than for a word with low base probability.

In this interpretation of the experimental data, word frequency serves as an estimate of base probability. In the cited experiments word frequency was determined from the published tables of the Lorge Magazine Count (10).* This count is based on a sample of 4.5 million words of text taken from issues of five popular magazines dating from the period 1928-1939. The principal experiment correlating duration threshold with word frequency, however, was carried out in the Summer of 1948 with Harvard undergraduates as subjects (3). Hence the question arises: to what extent can word frequencies based on the linguistic behavior of magazine writers in the 1930s represent the average base probabilities of Harvard students in 1948? We shall consider that question here as it applies to the 75 words used in the two main recognition experiments (3, Table 1 and p. 406).

PROCEDURE

To ascertain directly the degree of relationship between word frequency and base probability we would need to correlate Magazine-Count frequencies with frequencies obtained from a sample of comparable size taken from the total linguistic production of Harvard undergraduates in the Summer of 1948. Preparation of a count of student language on such a scale, however, is at present unfeasible. Correlation of Magazine-Count frequencies with frequencies based on a small sample of student language would be unsatisfactory because many of the words used in the recognition experiments have very small probabilities that could not be estimated from samples of less than a million words. We are therefore forced to rely upon an indirect technique of measuring the degree of relationship.

Three experiments are reported below. In each experiment, student subjects were asked to rank a set of words according to the frequency with which those words are used by their own college community. The rank of a word, averaged over all subjects, is assumed to estimate the relative base probability of the word for that population. In other words, it is assumed that a group of students asked which of two words they use more frequently will, more often than not, choose the word that in actual fact occurs more frequently. These student ranks were then correlated with ranks for the same words based on the Magazine-Count tables in order to obtain an estimate of the degree of correlation between relative base probability and relative Magazine-Count frequency. Strictly considered, the results of this type of test apply only to ranks, although we shall see later that under certain conditions they also apply to the frequencies themselves.

---

*In some of the experiments the Thorndike-Lorge Semantic Count was also used. Since the correlations with duration threshold are about the same whether the Semantic Count or the Magazine Count is used (3), only the latter will be considered here.

The studies to be described all depend on the same computational procedures. To obtain data representing the population of students as a whole, the ranks assigned each word by the different subjects were totalled and a set of average ranks computed for these totals. These average ranks were then correlated with Magazine-Count* rank by means of Spearman's coefficient ρ (6, p. 106). In order to correct these rank-order coefficients for attenuation, the reliabilities of the two sets of ranks had to be estimated. The reliability of the students' estimates was determined by a conventional split-half technique. The subjects were divided randomly into two subgroups and the words ranked for each subgroup just as they were for the total group. The reliability of the total group was then estimated from the rank-order correlation between the two subgroups by means of the Spearman-Brown prophecy formula (6, p. 194). The reliability of the ranks based on Magazine-Count frequency could not be measured by a true split-half technique, there being no way to divide the total count randomly into two subgroups. But the Magazine Count records separately the frequency of a word in each of the five different magazines sampled. Most of the systematic differences between magazines can be cancelled out by pooling three of them (The Saturday Evening Post, The Woman's Home Companion, and The Reader's Digest) into one subgroup of 2.38 million words and the other two (True Story and The Ladies' Home Journal) into a second subgroup of 2.21 million words. Rank-order correlations between the frequencies of the words in these two subgroups were used to obtain an estimate of the reliability of the Magazine-Count ranks. The two reliability coefficients were then used to correct for attenuation the correlation between average student ranks and Magazine-Count ranks.

EXPERIMENTS

Experiment 1. This experiment was carried out in 1948 using 14 Harvard College students as subjects. It is therefore directly applicable to the population used in the experiments on duration threshold. The words were the 25 rarest ones in the list of 60 used in the main threshold experiment (3, Table 1). It has generally been supposed (e.g., 1) that the frequencies of rare words are more apt to depend on peculiarities of the sample chosen for a word count than are the frequencies of common words. The correlation for the entire set of 60 words thus should be at least as high as the correlation for the 25 rarest ones.

The subjects' instructions were to "rank the words in order of their frequency of occurrence among Harvard undergraduates. Frequency of usage refers to occurrence in all forms of language — speaking and listening,

*The unpublished version, giving the frequencies for words in their fully-inflected forms, was used. I wish to thank Dr. Irving Lorge for permission to use this version of the count.

reading and writing. A slight emphasis on speaking frequencies is probably justified since speech is probably the most basic form of language. Use your own verbal behavior as typical of 'Harvard undergraduates' but take into account any very atypical idiosyncracies of your speech." The subjects were also told to consider the words in their specific grammatical forms, not as root words. Since the words were all taken from the lowest frequencies in the Magazine Count, there were a number of cases in which different words had the same Magazine-Count frequency. Wherever possible, such ties were broken by giving the higher rank to the word whose frequency is higher in the Thorndike-Lorge Semantic Count (10). This count is based on a sample of approximately the same size as the Magazine Count but taken from more literary sources, such as the Encyclopedia Britannica, the Literary Digest, and miscellaneous novels and textbooks.

Table 1 gives the Magazine-Count ranks and the average student

TABLE 1

Words of Experiment 1 Listed in Order of
Magazine-Count Rank with Their Average Student Ranks

(Brackets Inclose Words with Identical Magazine-Count Frequencies)

Word	Rank	Word	Rank
celestial	11	erudition	6.5
assiduous	8.5	conviviality	14.5
benign	8.5	etcher	23
altruistic	2	psychical	22
amicable	6.5	inductive	3
mundane	5	pedagogue	13
condolence	12	vignette	20
metaphor	1	theistic	19
frugality	10	statics	18
beatific	16	elegies	17
barrister	14.5	chancels	25
rebuttal	4	{uncoerced	21
		{percipience	24

ranks for the 25 words of Experiment 1. The correlation between them is .71, corrected for attenuation to .78. These correlations are well above the usual criteria for statistical significance (in testing significance ρ was interpreted as r). The reliabilities are also of interest: for the students' rankings the reliability ρ is .988; for the Magazine-Count ranks, .835.

Experiment 2. This study, also carried out with Harvard College students in 1948, was designed to test the upper limit of the students' ability to estimate base probabilities. To this end the 15 words used in a supplementary threshold experiment (3, p. 406) were selected. These words cover a large range of Magazine-Count frequency (o to 1500) and differ by approximately equal logarithmic distances so that no two words are in the same frequency range. High reliability was assured by including only words that have the same rank in both Magazine and Semantic Counts. Fourteen subjects were used (no subject was used in both Experiments 1 and 2). The instructions were to rank the words "on the basis of how common you think these words are in their usage by college students."

Magazine-Count ranks and average student ranks for these words appear in Table 2. The correlation is .87, corrected for attenuation

TABLE 2

Words of Experiment 2 Listed in Order of Magazine-Count Rank with Their Average Student Ranks

Word	Rank	Word	Rank
country	1.5	testify	7
promise	5	surmise	11
example	1.5	dwindle	9
balance	3	irksome	12
welfare	6	vulture	13
venture	8	machete	15
deserve	4	titular	14
		figment	10

to .88. The reliability of the students' ranks is .991; that of the Magazine-Count ranks, .983.

<u>Experiment 3</u>. A further study was carried out in 1953 on all 60 words of the main threshold experiment using Antioch College students as subjects. The 60 words were divided into three lists having approximately the same distributions of Magazine-Count frequencies. Each list was given to 10 subjects. The instructions were to "rank the words according to the frequency with which you think they are used in the Antioch Community. Frequency of usage here includes <u>all</u> occurrences of words, in both written and spoken language, during the present school year (1952-3). Consider the words exactly as they are written, not their roots or related forms."

The results are shown in Table 3. The correlations are .81, .82,

TABLE 3

Words of Experiment 3 Listed in Order of
Magazine-Count Rank with Their Average Student Ranks

(Brackets Inclose Words with Identical Magazine-Count Frequencies)

Part A Word	Part A Rank	Part B Word	Part B Rank	Part C Word	Part C Rank
picture	2	government	4	service	3
market	6	education	1.5	knowledge	2
friendly	1	sympathy	7	automobile	7.5
savings	8	{scientific	1.5	lawyer	9
spiritual	7	painting}	9	religious	5
hospitality	9	orchestra	10	churches	10
literary	3	economics	6	heavenly	11
ensemble	12.5	poetry	8	broker	14
assets	5	intellectual	3	limousine	17.5
{charitable	14	initiative	5	liberties	1
earthly}	11	benign	17	assiduous	15
reverence	12.5	amicable	12	physics	4
debating	4	mundane	11	condolence	12
celestial	17	{metaphor	13	{judiciary	7.5
altruistic	10	beatific}	19	frugality}	13
{uncoerced	15	{rebuttal	14	{erudition	16
statics	16	barrister}	20	etcher}	17.5
vignette	18	conviviality	18	{inductive	6
pedagogue	19	{theistic	15	percipience	19
chancels}	20	psychical}	16	elegies}	20

and .57, respectively, for the three parts of the experiment. Corrected for attenuation, these become .85, .84, and .60. All are statistically reliable. The reliabilities for the students' estimates are .950, .982, and .942; for the Magazine-Count ranks, .978, .950, and .952, respectively.

Application to relative base probabilities. These experiments give rank correlations between word frequency and student estimates of relative base probability. They can be interpreted as rank correlations between word frequency and base probability only on the assumption that the student estimates of relative base probability were correct. Some tests of that assumption are considered in this section.

In Experiments 2 and 3 each subject was asked to give a second set of rankings in which he ordered the words according to their frequencies in his own personal usage. The exact instructions in Experiment 2 were to rank the words "on the basis of your impression of how common the words are to you personally"; and in Experiment 3, to "rank the words according to the frequency with which you yourself use them." It will be convenient to refer to these as personal ranks and the ranks previously described, where the words were ranked according to their use by the entire college population, as college ranks. Comparison of the two sets of ranks permit us to test the ability of the students to rank base probabilities.

Let us observe how the two sets of student rankings are related. The base probability of a word, if we could measure it, would be found to vary somewhat from one student to another. If these personal base probabilities were averaged over all students in the defined student population, the result would give the base probability of the word for the college population. The assumption that the students were correct in their estimates of relative base probability then implies the following relations:

(1) Personal ranks should vary more from one subject to another than college ranks, since the former reflect real differences in base probability as well as errors of estimation while the latter reflect errors of estimation only. Hence the interpersonal correlations for personal ranks should be lower than the corresponding interpersonal correlations for college ranks.

(2) Personal ranks averaged over the entire population of subjects should (when a sufficiently large sample of the population is taken) equal the average college ranks.* From this it follows that Magazine-Count ranks should correlate as highly with average personal ranks as with average college ranks.

The first deduction can be tested by comparing the rank correlations for all possible combinations of subjects in Experiments 2 and 3. For each type of ranking there are a total of $\binom{14}{2} = 91$ correlations in Experiment 2 and a total of $\binom{10}{2} = 45$ correlations in each part of Experiment 3. Well over half the correlations are higher for the college ranks, the proportions being 57/91 in Experiment 2 and 25/45, 32/45, and 29/45 for the three parts of Experiment 3. The total proportion, 143/226, is four standard-error units above the ratio of 113/226 that would be expected if there were no difference between the

---

*This assumes that ranks according to frequency can be averaged like the actual frequencies, an assumption that is not generally valid. But the deduction here concerns only the relative magnitudes of ρ. Any bias introduced by the averaging of ranks should affect all ρ's alike, since by Zipf's law (12) the distributions of word frequency will be similar. Hence the conclusions should not be affected materially by the averaging of ranks.

two types of correlations. The superiority of the correlations for college ranks can also be shown by t-tests for the differences between mean correlations. The t's are 3.5 for Experiment 2 and 3.2, 4.3, and 3.8 for the three parts of Experiment 3, all significant at the .01 level.

According to the second deduction, the correlations between average personal ranks and average college ranks should approach unity after correction for attenuation. The raw correlations between average ranks are .967 for Experiment 1 and .964, .990, and .962 for the three parts of Experiment 3. Corrected for attenuation, they become .978, .998, 1.016, and 1.006, respectively. The root-mean-square of these coefficients, .9996, is remarkably close to the predicted value of 1.

It remains to see whether the Magazine-Count ranks correlate as highly with average personal ranks as with average college ranks. The respective rank correlations, both with and without correction for attenuation, are shown in Table 4. The root-mean-square correlation

TABLE 4

Correlations of Personal and of College Ranks With Magazine-Count Ranks, With and Without Correction for Attenuation

	Type of Correlation			
Experiment	Uncorrected		Corrected	
	Pers	Coll	Pers	Coll
2	.89	.87	.90	.88
3-A	.77	.81	.79	.85
3-B	.83	.82	.86	.84
3-C	.59	.57	.62	.60
Root-mean-squares	.78	.78	.80	.80

is .778 for average personal ranks, .775 for average college ranks. After correction for attenuation the root-mean-square correlations are .801 and .799, respectively. In both cases the differences are negligible.

The results show that personal ranks differ more from subject to subject than do college ranks, but that these individual differences in personal ranks cancel out in such a way as to approach the college ranks for groups of 10 subjects or more. This is the picture that would be expected on the assumption that the students' estimates of relative base probability were valid.

Application to base probabilities. The data reported above concern only the order of words with respect to their frequency of occurrence, and the conclusions have been phrased accordingly. It is possible to extend the conclusions to the actual frequencies, however, without violating any of the assumptions underlying the statistical procedure. The coefficient ρ measuring the correlation between two sets of ranks also measures the correlation between the variables ranked if those variables are distributed rectangularly (6, pp. 106f.). Now the logarithms of the Magazine-Count frequencies of the words used in each of the above experiments form approximately rectangular distributions (cf. 3, Fig. 1). The actual base probabilities of the words in the students' usage are not known, so the form of their distribution cannot be determined directly. But it is reasonable to suppose that their distributions do not differ materially in form from the Magazine-Count distributions in view of Zipf's evidence that the distribution of word frequencies has the same mathematical form for all heterogeneous samples of language (12). The rank correlations reported here may thus be regarded as close approximations to the product-moment correlations between log Magazine-Count frequency and log base probability.

On the basis of Experiments 1 and 3 the reliability of measurements of log base probability for the words of the main threshold experiment can be put at about .75. The raw product-moment correlation between log Magazine-Count frequency and mean duration threshold in that experiment is -.68. When the thresholds have been corrected for certain stimulus characteristics the raw correlation becomes -.76. The reliability of mean duration threshold in that experiment is about .90. Hence the correlation between duration threshold and log base probability, corrected for attenuation, can be estimated at about -.83 without the correction for stimulus characteristics or about -.93 with that correction. These indicate a high degree of rectilinear relationship between the two variables.

DISCUSSION

Other interpretations of the correlation between word frequency and duration threshold have been suggested. There are two main points in which these differ from the present interpretation. Usually word frequency is interpreted as the frequency with which a word has occurred in the past history of the subject rather than an estimate of the probability of the word at the time of recognition, and as the frequency with which the subjects see or read the word rather than the frequency with which they emit it as a response (5, 7, 8, 9, 11). Some of the reasons for rejecting these points of view in favor of the present interpretation are mentioned in this section.

According to the first alternate interpretation, Magazine-Count frequencies serve as estimates of the frequencies with which words have occurred in the past histories of 1948 Harvard undergraduates. If we considered only the students' linguistic histories immediately prior to the recognition experiments, the error of estimation introduced by differences between those histories and the material used for the Magazine-Count would be about the same as the error for the response-emission interpretation. But error of a greater order of magnitude must be expected when the Magazine Count, based on adult language, is used to represent the students' language during childhood and adolescence, which make up the major part of their total linguistic histories. It is not quite clear, moreover, just how a word's frequency of previous occurrence could have an appreciable effect upon its threshold. Almost all studies indicate that learning reaches an asymptote as a function of practice. Thus repetition of an event after a very large number of previous repetitions has no appreciable effect on behavior. Even rare words with frequencies of 5 or 10 in the Magazine Count must have occurred often enough among the total production of words in a student's life history to have reached the asymptote of learning. Hence the observed differences in duration threshold between words of these low frequencies and words of much higher frequencies can hardly be attributed to differences in the number of times the words have occurred in the subjects' pasts.

The second point to be considered is whether Magazine-Count frequency should be interpreted as an estimate of the frequency with which the subjects read a word or the frequency with which they emit it. Both the language read and the language emitted by Harvard undergraduates can be expected to differ significantly from the material of the Magazine Count. The reading-frequency interpretation thus involves the same kind of error that is estimated in the previous section for the response-frequency interpretation. But additional sources of error will affect the estimation of reading frequencies. In normal reading a person often skips connective words or words at the ends of lines when the meanings are indicated by context, and a difficult passage or unfamiliar word, on the other hand, may be read over and over. Two words occurring with the same frequency in a sample of reading material, consequently, may differ considerably with respect to the number of times they are actually read. Furthermore, it is difficult to specify the frequency with which a word is read except in terms of the time spent in reading it, since there is no simple observable response that can serve as a criterion of reading. But the amount of time spent in reading a word will be affected by many factors other than the number of times it appears in the material that is read. Hence the Magazine Count will give a far more erroneous estimate of the frequency with which a word is read by Harvard undergraduates than of the frequency with which they emit it.

Some recent experimental evidence also counters the reading-frequency interpretation. The frequency with which students actually read a word is better estimated from the frequency of the constituent sequence of letters, computed without regard to their occurrence as a complete word, than from the frequency of the complete word. Experiment, however, shows no correlation at all between letter-sequence frequencies and duration threshold (7, p. 76). While it can be argued that these results are not conclusive, since the letter-sequence frequencies were based on a much smaller sample than the word frequencies, the complete lack of correlation strongly suggests that the frequency with which a word is read cannot account for the correlations of .6 to .7 that have been found for word frequency and duration threshold.

SUMMARY

An interpretation of the inverse relationship between the duration threshold of a word and its frequency of occurrence is outlined. According to this interpretation, the frequency of a word in the Thorndike-Lorge tables (10) serves as an estimate of the frequency with which college students would have used that word at the time the duration thresholds were measured if the measurements had not been made. The validity of this estimate is tested by three experiments based on a rank-correlation procedure. Additional experiments provide a check on the method. The results indicate a validity of about .75 for the subjects used in the principal experiment on duration threshold. Some reasons for preferring the proposed interpretation to others that have been suggested are briefly mentioned.

REFERENCES

1. Dewey, G. The relative frequency of English speech sounds. Cambridge, Mass.: Harvard Univ. Press, 1923.

2. Howes, D.H., and Solomon, R.L. A note on McGinnies' "Emotionality and perceptual defense." Psychol. Rev., 1950, 57, 229-234.

3. Howes, D.H., and Solomon, R.L. Visual duration threshold as a function of word probability. J. exp. Psychol., 1951, 41, 401-410.

4. McGinnies, E.M., Comer, P.B., and Lacey, O.L. Visual recognition thresholds as a function of word length and word frequency. J. exp. Psychol., 1952, 44, 65-69.

5. Miller, G.A. Language and communication. New York: McGraw-Hill, 1951.

6. Peters, C.C., and Van Voorhis, W.R. Statistical procedures and their mathematical bases. New York: McGraw-Hill, 1940.

7. Postman, L. The experimental analysis of motivational factors in perception. In J.S. Brown, et al., Current theory and research in motivation: a symposium. Lincoln: Univ. of Nebraska Press, 1953.

8. Postman, L., and Schneider, B.H. Personal values, visual recognition, and recall. Psychol. Rev., 1951, 58, 271-284.

9. Solomon, R.L., and Postman, L. Frequency of usage as a determinant of recognition thresholds for words. J. exp. Psychol., 1952, 43, 195-201.

10. Thorndike, E.L., and Lorge, I.D.M. The teacher's word book of 30,000 words. New York: Teachers College, Columbia Univ., 1944.

11. Wispe, L.G., and Drambarean, N.C. Physiological need, word frequency, and visual duration thresholds. J. exp. Psychol., 1953, 46, 25-31.

12. Zipf, G.K. Human behavior and the principle of least effort. Cambridge, Mass.: Addison-Wesley Press, 1949.