Bulletin of the Psychonomic Society
1985, 23 (4), 377-380

Effects of instructions, orienting task, and memory tests on memory for words and word frequency

JAMES R. ERICKSON and CAROL RENAUD GAFFNEY
University of Texas, Arlington, Texas

Common words were presented to subjects who either rated them for pleasantness or estimated letter frequency. Words were presented with frequencies of 0, 1, 3, or 6; half of the subjects were instructed to attend to frequency, half were not. Frequency and memory tests were then given, with half of the subjects receiving each test type first. Memory was improved by semantic processing, greater presentation frequency, and additional exposure during frequency tests. Frequency judgments were accurate when tested first, but frequency discrimination declined drastically when frequency was tested second. Semantic processing produced more accurate relative frequency judgments. Data are discussed in terms of models of frequency judgment, and artifacts of different kinds of data analysis are noted.

Psychologists have been interested in frequency for many years. Often the interest has been in *effects* of frequency—repetitions—on perception, learning, memory, etc. Recently, however, attention has been paid to frequency per se. There is now substantial literature attesting to the accuracy with which absolute and relative frequency judgments are made, both in laboratory (e.g., Hintzman & Block, 1971; Howell, 1973b) and in naturalistic (e.g., Attneave, 1953; Lichtenstein, Slovic, Fischhoff, Layman, & Combs, 1978) settings. Several theorists have considered ways in which frequencies come to be represented in memory (e.g., Hintzman, 1976; Howell, 1973a; Whitlow & Estes, 1979); others have discussed heuristics, such as availability, that could underlie frequency judgments (e.g., Combs & Slovic, 1979; Tversky & Kahneman, 1973).

One of the more intriguing theoretical positions is that humans essentially are "hard wired" to track frequencies automatically. Although it is not difficult to carry this idea to a reductio-ad-absurdum by considering finer and finer grain stimulus analysis, Hasher and Zacks (1979; Zacks, Hasher, & Sanft, 1982) have presented some impressive data supporting this general view, not only for frequencies of nominal stimuli presented to subjects, but also for various properties of such stimuli (e.g., Alba, Chromiak, Hasher, & Attig, 1980; Gude & Zechmeister, 1975).

One kind of data supporting the notion of automatic processing of frequencies is evidence that frequency judgments are usually no more accurate when people are instructed to attend to frequencies than when they are not (Flexser & Bower, 1975; Howell, 1973b; Zacks et al., 1982). Related to this is the finding that subjects who are instructed that they will have to recall words and then are given a surprise frequency judgment task show no decrement in frequency judgments, but subjects who are instructed to attend to frequencies and then are given an unexpected recall test show poorer recall than those given accurate instructions (Howell, 1973b; Zacks et al., 1982). Indeed, tests for recall and for frequency often produce rather different effects, indicating that the representations of stimuli and of their frequencies may be relatively independent (e.g., Underwood, Zimmerman, & Freund, 1971).

There is one task manipulation that has produced rather consistent effects on frequency judgment. Subjects who are asked to make a semantic judgment about stimulus words as they are presented (see Hyde & Jenkins, 1973, for examples) have been consistently more accurate on later frequency judgments than those asked to make a nonsemantic judgment (Fisk & Schneider, 1984; Rose & Rowe, 1976; Rowe, 1974). This effect, of course, is not predicted by automatic processing theories of frequency information.

In the present study, the two variables noted above were manipulated simultaneously. Subjects either were or were not instructed to attend to the frequencies with which words were presented, and subjects were asked to make either a semantic (pleasantness) or a nonsemantic (number of letters) rating of each stimulus word at presentation. They were then tested for frequency knowledge and for memory of the presented words. Although there are several articles in which recall and frequency judgments have been compared in between-subject designs, we have not encountered one in which subjects were asked to produce both kinds of information. Thus the extent to which frequency and item memories are based on different kinds of information, in which case memory tests would be relatively independent, or on the same or similar representations, in which case one kind of memory

The authors' mailing address is: Department of Psychology, Room 313 Life Science Buildings, Box 19528, Arlington, TX 76019-0528.

test could influence another, is not known. Because this type of study is potentially important theoretically, it seemed useful to explore effects of item tests on frequency judgments and vice versa. Half of our subjects made frequency judgments first, and then were tested for recall and recognition, and half received the tests in the opposite order.

METHOD

Subjects

Subjects were 80 students from the UTA introductory psychology subject pool. These students can either volunteer for particular experiments or write reports on published experiments to satisfy a course requirement.

Apparatus and Materials

All stimulus materials used were common words, chosen from Category 8 of the Toglia and Battig (1978) norms. Words which were obvious close associates of one another were not used. With this restriction, 86 words were chosen with the use of a random number table (RAND, 1955). Forty words were "old" items, presented in random order with frequencies of either 0, 1, 3, or 6 during the training phase of the experiment. Forty other words were "new" words, used during a recognition memory test. The other 6 words were fillers; 3 were presented at the beginning and 3 at the end of the training list, but they were not later tested.

The resulting 106-item training list (10 words presented six times; 10, three times; 10 once; and 6 filler items) was tape recorded, at a rate of 4 sec per word, for presentation to subjects. As the tape was played, half the subjects responded to each word by rating it for pleasantness and half responded by estimating the number of letters in it. Words varied in length from 3 to 10 letters; subjects were asked not to count letters, but to estimate. Ratings were made on numbered answer sheets which had either five-point rating scales or blank spaces beside each number.

Design and Procedure

Subjects signed up for one of several experimental sessions, which were assigned at random to one of the groups formed by crossing rating task (pleasantness rating or letter estimation) and frequency instructions (half of the subjects were told to try to learn how often each word they rated was presented, and half were told nothing about any test to follow). After hearing and rating the words on the training list, subjects were given test booklets. Half of the subjects first rated each of the 40 "old" words for frequency (they were not told that 10 of these words had not been presented previously). They were then asked to make relative frequency judgments for the old words on a second sheet. The 10 words within a given frequency category were paired with words from all categories. For example, the 10 words presented six times were randomly divided as follows: 2 were paired with zero-frequency words, 2 with words presented once, 2 with words presented three times, and 2 with other words presented six times. Subjects were asked to guess if they were not sure which word of a given pair was presented more often.

Then two memory tests were given. To test recall, we gave subjects a page numbered from 1 to 40 and asked them to write down as many different words as they could from the list that had been read to them. Finally, a two-alternative forced-choice recognition test was administered. Each of the 40 old words was randomly paired with a new word, and subjects were asked to check the word from each pair that had been read to them.

The other half of the subjects received the recall and recognition tests first, followed by the frequency tests. The time allowed for the various tests was as follows: recall, 5 min; recognition, 2 min; absolute frequency judgment, 3 min; relative frequency judgment, 2 min.

Table 1
Probability of Recall and Estimated Frequency as a Function of Presentation Frequency, Orienting Task, and Test Order

		Mean Recall Probability				Mean Estimated Frequency			
		Frequency				Frequency			
Order	Orienting Task	0	1	3	6	0	1	3	6
Tested First	Pleasantness Rating	.00	.20	.60	.78	.51	1.36	3.57	6.54
	Letter Frequency	.00	.14	.46	.62	.52	1.76	4.42	7.12
Tested Second	Pleasantness Rating	.13	.60	.76	.83	1.14	1.56	3.37	5.14
	Letter Frequency	.07	.38	.64	.68	1.18	1.94	3.42	4.89

Note—n = 200 per data point (10 subjects × 10 items × 2 instructions).

RESULTS

Recall

The recall data are shown in Table 1, which gives the probability of recall for old words as a function of presentation frequency, rating task, and test order. An ANOVA was performed using the number of words recalled from frequency categories 1, 3, and 6 (the words actually presented). The main effects of the following variables were significant: frequency [$F(2,144) = 180.50$], rating task [$F(1,72) = 40.63$], and test order [$F(1,72) = 66.48$] (all ps $< .001$; MS subjects/groups = 3.03; MS frequency × subjects/groups = 1.81). The interaction of frequency and test order was also significant [$F(2,144) = 18.07$, $p < .001$], and the interaction of frequency instructions and test order was marginally significant [$F(1,72) = 3.43$, $p < .07$]. A test of simple effects for this interaction showed that when the recall test was given first, recall was poorer [mean p(recall) = .43 vs. .51; $F(1,72) = 10.63$, $p < .01$] if subjects had been asked to attend to frequencies than if they had not. When the test for recall was given second, the effect of frequency instructions was essentially zero.

The significant effects are all obvious in Table 1. Recall probability increased monotonically with presentation frequency, was higher when the rating task was semantic, and was higher when the recall test was given after frequency judgments had been made. All of these effects were expected. Taking the frequency tests first gave subjects two additional exposures to the presented words (and their first exposure to words presented with zero frequency) and this additional exposure produced better recall. The frequency × order interaction probably reflects a ceiling effect; the "boost" in recall was greatest for words presented once during training, and declined monotonically with presentation frequency.

Recognition

Recognition performance was excellent; in fact, there were too few recognition errors to allow an analysis by presentation frequency. (Descriptively, there were fewer errors for words presented six times than for those presented one or three times.) For each subject, errors were combined over presentation frequencies of 1, 3, and 6, and these data were submitted to an ANOVA. The main effects of rating task [$F(1,72) = 8.07$] and test order [$F(1,72) = 8.07$], and their interaction [$F(1,72) = 11.27$],

were all significant (all ps < .01; MS subjects/groups = 0.75). The data for these effects are shown in probability form in Table 2. All the significant effects seem to be due to the fact that recognition performance was poorest when the memory tests were given first and the rating task was nonsemantic. All other cell means are essentially equal. No other effects approached significance. The recognition data are similar to the recall data, expect for the fact that there was no detectable effect of frequency instructions on recognition performance.

Absolute Frequency Judgment

Mean absolute frequency judgments are shown in Table 1 as a function of presentation frequency, test order, and rating task. An ANOVA including all frequency categories showed that the main effects of test order [$F(1,72) = 4.07$, $p < .05$], of presentation frequency [$F(3,216) = 387.02$, $p < .001$], and their interaction [$F(3,216) = 21.81$, $p < .001$] were all significant (MS subjects/groups = 3.04; MS frequency × subjects/groups = 1.07). No other effects or interactions approached significance. The frequency estimation functions for the two test orders were essentially linear, but with different slopes and intercepts. The regression equations of mean judged frequency on actual frequency were:

Frequency test first: Judged Frequency = .58 + 1.06 (Actual Frequency)

Frequency test second: Judged Frequency = 1.19 + .65 (Actual Frequency)

The first equation shows very accurate frequency judgments, with a positive bias of about half a unit. Possibly this bias would have been reduced if subjects had been told that some items on the test list had not been presented during training. When the frequency tests were given second, frequency performance declined considerably in accuracy. The bias increased, and the slope of the function declined. It might be noted that the index suggested by Flexser and Bower (1975), namely the correlation between judged and actual frequency, was .99 for both test orders. There is better information in the slopes and intercepts of the regression functions than in the correlation coefficients, and the Flexser-Bower index cannot be recommended by itself. When the regression functions are really linear, the correlations must be close to 1 and can provide no differential information. The decrease in the slope of the frequency estimation function when the frequency test was preceded by memory tests is discussed in more detail below.

Relative Frequency Judgment

As with the recognition data, there were too few errors on the relative frequency test to allow an analysis by presentation frequency. (Descriptively, most of the errors that did occur were on the 6 vs. 3, 1 vs. 0, and 3 vs. 1 items). Therefore the number of errors was tallied for each subject for items where the presentation frequencies were different, and these data were submitted to an ANOVA. These data are shown in probability form in Table 2 as a function of test order and rating task. Only the effect of rating task was significant [$F(1,72) = 7.39$, $p < .01$, MS subjects/groups = 0.74]. As can be seen in Table 2, subjects were more accurate when they had processed the words semantically during training. This replicates the orienting task effect found in several other studies (Fisk & Schneider, 1984; Rose & Rowe, 1976; Rowe, 1974).

Indirect Measures of Recognition

It is reasonably common in frequency judgment experiments to use the frequency judgment data to estimate recognition memory (e.g., Underwood et al., 1971). For example, presented items with judged frequencies greater than zero might be labeled recognition "hits." Data from the present experiment were so analyzed and are shown in Table 2 along with the actual recognition data. The data from the relative frequency tests are from items where a word presented one, three, or six times was compared with one not presented, and should provide a reasonably direct analog of the forced-choice recognition test. However, as can be seen in Table 2, the measures are

Table 2
Relative Frequency Judgments and Actual and Estimated Recognition Performance as a Function of Orienting Task and Test Order

		Relative Frequency Judgments	Actual Recognition Performance	Estimated Recognition Performance		
				From Relative Frequency	From Absolute Frequency	
Order	Orienting Task	p(Correct)	p(Correct)	p(Correct)	p(Hit)	p(False Alarm)
Tested First	Pleasantness Rating	.950*	.992†	.958‡	.992†	.346§
	Letter Frequency	.883	.952	.900	.943	.337
Tested Second	Pleasantness Rating	.908	.988	.908	.997	.944
	Letter Frequency	.88	.992	.892	.992	.810

NOTES—**n = 240 per data point (10 subjects × 12 items × 2 instructions). †n = 600 per data point (10 subjects × 30 items × 2 instructions). ‡n = 120 per data point (10 subjects × 6 items × 2 instructions). §n = 200 per data point (10 subjects × 10 items × 2 instructions).*

not very similar. Recognition performance estimated from frequency judgments was not as high as actual recognition performance, and the ordinal relations among conditions were not the same for the two measures.

The hit rates from the absolute frequency ratings (the probability that an item presented one, three, or six times received a frequency rating of one or more) do seem to match the recognition data well. But, of course, judgments of absolute frequencies are subject to bias, and the false alarm data must also be considered. The false alarm rates (probably that an item not presented received a frequency rating of one or more) were quite high, particularly when the frequency tests were given second. A more appropriate performance measure for comparison, such as d′ from signal detection theory, shows large differences between the actual recognition performance and that estimated from absolute frequency judgments.

It should be obvious that one should interpret indirect measures of recognition with a good deal of caution, if not with a grain of salt. Subjects may not approach a frequency judgment task with the same "set" as a recognition task, or may not use the same kind of information to make their judgments.

DISCUSSION

The recall and recognition data are quite straightforward and replicate well known effects. Repeating a word (at least in spaced fashion) during training produces better memory for that word. Attending to a word semantically improves recall. Instructions to attend to the frequency with which words are presented interfere with recall of the word. And, of course, additional exposure to words (as occurred during the frequency tests) produces better performance on memory tests. None of these effects is surprising, and the memory data require no further discussion.

The effects of the orienting task were as expected: Attending to the meaning of a word as it is presented seems to provide better access to that word in a variety of memory tasks. In the present experiment, relative frequency judgments were more accurate for subjects who had rated words for pleasantness at presentation. The absolute frequency ratings for these subjects were also more accurate, but not significantly so. This orientation effect replicates results of other researchers and remains to be explained by automatic processing theories of frequency information.

There was one surprising effect in the frequency judgment data: the dramatic decrease in the accuracy of absolute frequency judgments when recall and recognition tests preceded frequency judgments. Prior data and intuition had led us to believe that the effects of prior memory tests would either be quite small, or that there would be increases in both the intercept and the slope of the frequency judgment function. The data of Hintzman and Block (1971, Experiment III) suggest that subjects can fairly accurately discriminate the frequencies with which words appear in two contexts (lists in their experiment). Because the memory test context would seem to be quite different from the training list context, one might thus expect little interference and quite similar frequency judgment data either preceding or following a test for word memory.

On the other hand, one might expect frequency judgments to be based on total situational frequency. In that case, the recognition tests provided an increase in situational frequency of exactly 1 for every old word, which could produce an increase in the intercept of the frequency judgment function. And because recall was so much better for words presented with high frequency (and was zero for words not presented at all), high frequency words should receive a greater increment in situational frequency than low frequency words receive, leading to an increase in the slope of the frequency judgment function.

In fact the intercept increased, but the slope decreased. The data are rather similar to those of Underwood et al. (1971), who examined frequency judgments after various test delays of up to 1 week. It is apparent that the memory tasks interfered with frequency discrimination. This interference was most obvious in the absolute judgments, but also appeared in relative judgments, although the decline was not significant in the latter case. The interference produced by prior memory tests was substantial and too large to be artifactual [e.g., due to the 7-min (differential) delay when frequency was tested second]. The apparent fragility of the representation of episodic frequency found in this experiment needs to be explored in future research.

REFERENCES

ALBA, J. W., CHROMIAK, W., HASHER, L., & ATTIG, M. S. (1980). Automatic encoding of category size information. *Journal of Experimental Psychology: Human Learning & Memory*, **6**, 370-378.

ATTNEAVE, F. (1953). Psychological probability as a function of experienced frequency. *Journal of Experimental Psychology*, **46**, 81-86.

COMBS, B., & SLOVIC, P. (1979). Causes of death: Biased newspaper coverage and biased judgments. *Journalism Quarterly*, **56**, 837-843.

FISK, A. D., & SCHNEIDER, W. (1984). Memory as a function of attention, level of processing, and automatization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 181-197.

FLEXSER, A. J., & BOWER, G. H. (1975). Further evidence regarding instructional effects on frequency judgments. *Bulletin of the Psychonomic Society*, **6**, 321-324.

GUDE, C., & ZECHMEISTER, E. B. (1975). Frequency judgments for "gist" of sentence meaning. *American Journal of Psychology*, **88**, 385-396.

HASHER, L., & ZACKS, R. T. (1979). Automatic and effortful process in memory. *Journal of Experimental Psychology: General*, **108**, 356-388.

HINTZMAN, D. L. (1976). Repetition and memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 10). New York: Academic Press.

HINTZMAN, D. L., & BLOCK, R. A. (1971). Repetition and memory: Evidence for a multiple trace hypothesis. *Journal of Experimental Psychology*, **88**, 297-306.

HOWELL, W. C. (1973a). Representation of frequency in memory. *Psychological Bulletin*, **80**, 44-53.

HOWELL, W. C. (1973b). Storage of events and event frequencies: A comparison of two paradigms in memory. *Journal of Experimental Psychology*, **98**, 260-263.

HYDE, T. S., & JENKINS, J. J. (1973). Recall for words as a function of semantic, graphic, and syntactic orienting tasks. *Journal of Verbal Learning & Verbal Behavior*, **12**, 471-480.

LICHTENSTEIN, S., SLOVIC, P., FISCHHOFF, B., LAYMAN, M., & COMBS, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning & Memory*, **4**, 551-578.

RAND CORPORATION (1955). *A million random digits*. New York: Free Press.

ROSE, R. J., & ROWE, E. J. (1976). Effects of orienting task and spacing of repetitions on frequency judgments. *Journal of Experimental Psychology: Human Learning & Memory*, **2**, 142-152.

ROWE, E. J. (1974). Depth of processing in a frequency judgment task. *Journal of Verbal Learning & Verbal Behavior*, **13**, 638-643.

TOGLIA, M. P., & BATTIG, W. F. (1978). *Handbook of semantic word norms*. Hillside, NJ: Erlbaum.

TVERSKY, A., & KAHNEMAN, D. (1973). Availability: A heuristic for judging frequencies. *Cognitive Psychology*, **4**, 207-232.

UNDERWOOD, B. J., ZIMMERMAN, J., & FREUND, J. S. (1971). Retention of frequency information with observations on recognition and recall. *Journal of Experimental Psychology*, **87**, 149-162.

WHITLOW, J. W., JR., & ESTES, W. K. (1979). Judgments of relative frequency in relation to shifts in event frequency: Evidence for a limited capacity model. *Journal of Experimental Psychology: Human Learning & Memory*, **5**, 395-408.

ZACKS, R. T., HASHER, L., & SANFT, H. (1982). Automatic encoding of event frequency: Further findings. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **97**,106-116.

(Manuscript received for publication April 12, 1985.)