Bulletin of the Psychonomic Society
1975, Vol. 6 (3), 321-324

Further evidence regarding instructional effects on frequency judgments

ARTHUR J. FLEXSER and GORDON H. BOWER
Stanford University, Stanford, California 94305

Previous experimental reports have provided contradictory evidence regarding instructional effects on frequency judgments. An experiment was performed to clarify these findings, in which frequency discrimination was compared for two groups of subjects. One group was instructed as to the nature of the forthcoming frequency judgment task, while the other was told to prepare for an unspecified memory task. To avoid the possibility of response bias effects, the frequency discrimination coefficient, a correlational measure, was used to assess performance. Howell's 1973 finding of no instructional effects on frequency judgments was replicated. An attempt is made to reconcile this result with the finding by Begg and Rowe that subjects in a continuous frequency judgment task gave unusually accurate mean frequency judgments compared to subjects who were tested following study of all items.

Interest in how information is structured in memory, and in the nature of the memory trace, has led in recent years to investigations of how humans are able to discriminate in memory between the frequencies of different events. A review by Howell (1973a) distinguishes four hypotheses regarding the mechanism underlying this ability: trace strength, multiple trace, multiple process, and numerical inference.

The trace-strength hypothesis holds that each repetition of an event strengthens a generic representation of that event in memory, and that a frequency judgment is derived from the current strength. The simplest version of this hypothesis maintains that this same trace strength also determines the level of performance in other tasks which rely upon the same memory trace, such as recall or recognition. However, experimental evidence has failed to show the expected strict correspondence between mean frequency judgments and probability of recall or recognition. For example, frequency judgments and recall are affected differently by instructions (Howell, 1973b) and primacy in serial order (Underwood, 1969b), and frequency judgments differ from recognition in respect to the effects of semantic context (Rowe, 1973). Wells (1974) showed, in addition, that the function relating recognition probability to mean frequency judgments is not independent of spacing or true frequency, a finding which is inconsistent with the simple version of the trace-strength hypothesis. A more sophisticated version of the trace-strength hypothesis, which does not necessarily require a close correspondence between frequency judgments and recognition or recall measures, has been proposed by Underwood (1969a). Underwood proposes that frequency is represented as just one of a set of attributes comprising the generic memory trace of a repeated event. This hypothesis differs from the simple strength hypothesis described above in that repetitions are assumed to boost the "strength" of the frequency attribute, rather than the strength of the memory trace as a whole.[1]

Opposed to these trace-strength views is the multiple-trace hypothesis, which holds that repetitions of an event result in separate memory traces rather than in the strengthening of a single trace. Within this framework, frequency estimates are seen as being derived from a count of the number of relevant memory traces accessed by a given probe. The result of such a count need not coincide exactly with the subject's frequency estimate; he may also use availability heuristics similar to those described by Tversky and Kahneman (1973) to mediate between the retrieval stage and the response stage of the frequency estimation process. Thus, the subject may sample only a fraction of the total number of memory traces pertaining to a repeated event, and make a frequency estimate based on how readily such instances come to mind.

The experimental evidence appears to favor the multiple-trace representation of frequency information over the trace-strength hypothesis (see Howell, 1973a). For example, Hintzman and Block (1971) presented the same target words with different frequencies in both of two experimental lists, and later required subjects to judge both List 1 and List 2 frequencies for each item. Their subjects were able to perform this task with remarkable accuracy, and their judgments were only slightly affected by frequency within the irrelevant list. This finding seems incompatible with either version of the trace-strength hypothesis, unless unwieldy complicating assumptions are added (see Jacoby, 1972).

This research was supported by Grant MH-13950-07 from the National Institute of Mental Health to Gordon H. Bower. Requests for reprints should be sent to Gordon H. Bower, Department of Psychology, Stanford University, Stanford, California 94305.

Studies such as that of Hintzman and Block do not, however, rule out the possibility that trace-strength and multiple-trace mechanisms may both account in varying degrees for the ability to discriminate event frequencies. Such a compromise, which Howell (1973a) refers to as the multiple-process hypothesis, is consistent with available experimental evidence, although not as parsimonious as the multiple-trace hypothesis.

The experiment to be reported here does not attempt to distinguish between the various mechanisms discussed above. Instead, our aim is to assess the plausibility of a fourth scheme, which Howell has named the "numerical inference" hypothesis. This final hypothesis holds that the subject simply keeps a running count of the number of times an event has occurred, and that he later recalls this count when a frequency estimate is requested. The numerical inference hypothesis supposes that the usual laboratory procedure of having the subject keep track of the frequency of a number of items within a list is effectively transformed by him into a paired associate task. It is assumed that each time the subject sees an item, he retrieves the numerical count most recently associated with that item (or assigns a count of zero if the item is not recognized), increments this count by one, and associates the new count with the item.

The numerical inference hypothesis is similar to the previously described frequency-attribute hypothesis in that both regard frequency as a directly stored rather than a derived attribute. That is, there is assumed to be a unique representation of frequency information contained in the memory trace(s) pertaining to a repeated event. When a frequency estimate is called for, the subject presumably accesses this information directly and responds on the basis of it. Such hypotheses may be contrasted to those which propose that a frequency estimate is *derived* from aspects of the event-memory representations which are not specifically oriented toward recording event frequencies.

Investigators have given scant attention to the numerical inference hypothesis. A probable reason is that the scheme is a deliberate artifice, whereas we are more interested in how humans arrive at frequency estimates in a natural setting than in special strategies which they may adopt only when they are in memory experiments. However, if subjects do employ this counting strategy in experimental frequency tasks, we would have to reinterpret earlier results in this light. Therefore, we should assess the extent to which this strategy is used in memory experiments.

A counting strategy seems to require voluntary effort on the subject's part. If so, then frequency discrimination should be worse for uninstructed subjects than for those who have been instructed in advance about the task. Two earlier experiments have investigated the effects of instructions upon frequency judgments, but their results have been inconsistent. Howell (1973b) gave subjects either frequency or recall instructions, and tested half of each class of subjects for either frequency judgments or free recall. Although recall instructions facilitated recall, there was no effect of instructional set on frequency judgments. The second study, by Rowe and Rose (Note 1), used three instructional sets: intentional (frequency instructions), nonspecific (instructions to prepare for an unspecified memory task), and incidental (instructions to rate the potency or "strength" aroused by each word). Contrary to Howell's finding, Rowe and Rose found that all three instructional sets differed significantly from one another: the average frequency judgments were highest in the incidental condition and lowest in the nonspecific condition. Intentional instructions produced frequency judgments intermediate in size between the two uninformed conditions.

These last results are difficult to interpret, however, because the observed differences between conditions could reflect either true differences in frequency discriminability or merely "response criterion" shifts such as are commonly found in recognition experiments. For example, the multiple-trace theorist might argue that incidental instructions bias subjects to incorporate borderline event recollections into their frequency count. Reporting only mean frequency judgments corresponding to each true frequency (as Rowe and Rose did) is analogous to reporting only hit rate in a recognition experiment; it is subject to bias effects. What is needed is a measure of subjects' ability to distinguish one frequency from another, analogous to the d' measure of signal detection theory. In analyzing the experiment which follows, we have adopted such a measure, the *discrimination coefficient.*

The discrimination coefficient is defined as the correlation coefficient between the true and judged values of a quantity (such as frequency) about which absolute judgments are made. The use of discrimination coefficients is appropriate for frequency judgments due to the fact that the relationship between true and judged frequency approximates a linear function, even after delay (Underwood, Zimmerman, & Freund, 1971). The slight negative acceleration which is typically found in frequency curves (e.g., Hintzman, 1969) does not seem serious enough to invalidate a measure of accuracy based on correlation coefficients. If a discrimination coefficient is calculated separately for each subject, a measure of variability is obtained, and statistical tests between conditions are possible using standard techniques.[2] It should be pointed out that the discrimination coefficient is not a measure of *absolute* accuracy: if the subject estimates a frequency for every item which is twice the actual frequency, he

Table 1
Mean Frequency Judgments in Two Instructional Conditions

Instructional Condition		Frequency							
		0	1	2	3	4	5	6	Mean
Task-instructed	Mean	.18	.95	2.12	2.95	3.44	4.30	4.92	2.58
	SEM	.08	.11	.31	.14	.20	.37	.25	.15
Uninstructed	Mean	.08	1.27	2.43	3.57	3.96	4.76	4.78	2.86
	SEM	.06	.13	.41	.24	.43	.58	.37	.26

Note–Standard errors are based on variance of subject means.

will still obtain a discrimination coefficient of 1.00. Rather, the discrimination coefficient measures how well the subject's responses distinguish one frequency from another—a measurement which is of fundamental theoretical interest.

Besides the previously mentioned inconclusive findings regarding instructional effects on frequency judgments, there is another finding which is consistent with the numerical inference hypothesis. Begg and Rowe (1972) employed a continuous frequency estimation procedure in which subjects were required to give a frequency estimate on every presentation of an item. Mean frequency estimates under these conditions were extraordinarily accurate; the usual overestimation of low frequencies and underestimation of high frequencies was absent. A follow-up study by Begg (1974) established that this accuracy depended upon the fact that subjects gave an incrementing frequency estimate each time an item was presented, and was not merely a consequence of the lack of a delay between study and test. A reasonable interpretation of this result (and essentially the one cited by Begg and Rowe) is that the continuous frequency judgment procedure encouraged subjects to adopt the counting strategy. These continuous frequency judgment studies thus suggest that the numerical inference strategy may play an important role in frequency estimation, at least under conditions which appear to maximize its benefits.

In the experiment reported below, we attempt to shed more light on the question of instructional effects on frequency judgments, using an analysis based on discrimination coefficients in order to eliminate possible response bias effects.

METHOD

Subjects

Twenty-four subjects, 9 females and 15 males, participated. The subjects were either recruited from the Palo Alto area by advertisement and were paid for their participation, or received course credit in an elementary psychology course at Stanford University.

Materials and Apparatus

The stimuli were nonsense syllables (consonant-vowel-consonants) selected on the basis of high meaningfulness from the Krueger list reprinted in Underwood and Schulz (1960). The syllables were also chosen to minimize acoustic and visual similarities between different stimuli. During the study phase, materials were presented on slides by a Kodak Carousel projector with presentation rate controlled by an electronic timer.

Design

All subjects studied and were tested on the same list, which consisted of 117 stimulus items. Items with token frequencies of 1 through 6 occurred in the list, with six exemplars of frequencies 1-3 and five exemplars of frequencies 4-6. Repetitions of a given item were always separated by 4-7 intervening items. Three buffer items (not tested) were included at the beginning and end of the list.

Procedure

Each subject was tested individually, and the first 20 subjects were divided equally between the two experimental conditions, which differed only in the contents of an instruction sheet given to the subject at the beginning of the experiment. The instructions for both conditions noted that the subject was to be shown a series of nonsense syllables, some of which would be repeated. In the task-instructed condition, the instructions went on to describe the frequency judgment task for which the subject was to prepare himself, while in the uninstructed condition, he was merely told to attend carefully to each item in preparation for a memory task which would be explained later.

The stimulus items were then projected at a 6-sec rate. Following the final slide, the subject was given a second instruction sheet describing the frequency judgment task. (For the task-instructed subjects, this sheet was largely a repetition of the earlier instructions.) Subjects were explicitly instructed to assign a frequency of 0 to test items which had not appeared earlier. The test form contained 39 items, consisting of the 33 study items plus 6 new items.

Following the test, subjects in the uninstructed group were asked the question, "Did you have any idea that the task might be judgment of frequency before you were given the test instructions?" Four subjects answered this question in the affirmative. Data from these subjects were not analyzed; instead, additional subjects were tested in the uninstructed condition until data were available from 10 uninstructed subjects who claimed not to have expected the forthcoming frequency judgment task.

RESULTS

Table 1 shows the mean judged frequency corresponding to each actual value of frequency for the two instructional conditions. The tendency for the uninstructed subjects to give slightly higher frequency judgments was not significant, $t(18) = .98$. (This nonsignificant difference was in the opposite direction from the difference found between the intentional and nonspecific instructional groups in the Rowe and Rose experiment.) A frequency discrimination coefficient (defined earlier as the correlation between true and

judged frequency) was calculated for each subject from the 39 test items; the mean of these discrimination coefficients was .78 for the task-instructed subjects and .75 for the uninstructed subjects. The difference between the two instructional conditions was not significant, $t(18) = .81$. Our results for this experiment are thus in agreement with Howell (1973b) in finding no effect of explicit frequency instructions upon later frequency judgments.

DISCUSSION

In view of the absence of instructional effects found in our experiment, we are puzzled by the finding of Begg (1974) and Begg and Rowe (1972) that subjects who made continuous frequency judgments as they progressed through a study list showed enhanced accuracy in their mean judgments. One would suppose that a task-informed subject in our experiment would naturally make a spontaneous frequency judgment to himself each time a word was presented, and would therefore be expected to show the same benefits therefrom as did Begg's subjects. Indeed, a number of subjects in the intentional condition, questioned informally after the experiment, claimed they tried to keep a running count for each item. A possible reconciliation of the seemingly discrepant findings is suggested in the further introspections of some of these task-informed subjects: several volunteered that they soon gave up counting because they found it hopeless trying to keep track of so many stimuli (39 types and 117 tokens) and their associated numbers. It is possible, therefore, that the difference in findings between our experiment and Begg's is simply due to the fact that his procedure forced subjects to persist in a beneficial strategy which they would otherwise have quickly abandoned in discouragement.

A further possibility is that the seeming superiority of frequency discrimination for subjects who made running frequency estimates in Begg's experiments may be more apparent than real. As discussed previously, the possibility of response bias effects cannot be eliminated when only mean frequency judgments are reported, as was the case in the Begg studies. Thus, it may be that making running judgments does not enhance frequency discrimination, but merely eliminates a bias which subjects have against giving frequency judgments which are extreme in either direction, relative to the task mean. According to this view, the function of the running judgments would be to give the subject a more accurate perception of the total distribution of true frequencies, rather than to enhance frequency discrimination for individual items. In this case, more accurate mean judgments would be obtained at the expense of larger variances within each frequency category.

In any event, our main points are twofold. First, in order to compare frequency discrimination across different experimental conditions, a measure such as the discrimination coefficient should be used in order to distinguish true differences in discriminability from response bias effects. Second, when such a measure is used, no differences are found between task-instructed and uninstructed subjects, indicating that special strategies adopted in response to task demands do not substantially aid frequency discrimination.

REFERENCE NOTE

1. Rowe, E. J., & Rose, R. J. *Instructional and spacing effects in judgment of frequency.* Paper presented at the meeting of the Canadian Psychological Association, Windsor, Ontario, June 1974.

REFERENCES

Begg, I. Estimation of word frequency in continuous and discrete tasks. *Journal of Experimental Psychology*, 1974, **102**, 1046-1052.

Begg, I., & Rowe, E. J. Continuous judgments of word frequency and familiarity. *Journal of Experimental Psychology*, 1972, **95**, 48-54.

Hays, W. L. *Statistics*. New York: Holt, Rinehart, and Winston, 1963.

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

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

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

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

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

Rowe, E. J. Frequency judgments and recognition of homonyms. *Journal of Verbal Learning and Verbal Behavior*, 1973, **12**, 440-447.

Tversky, A., & Kahneman, D. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 1973, **5**, 207-232.

Underwood, B. J. Attributes of memory. *Psychological Review*, 1969, **76**, 559-574. (a)

Underwood, B. J. Some correlates of item repetition in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, 1969, **8**, 83-94. (b)

Underwood, B. J., & Schulz, R. W. *Meaningfulness and verbal learning*. New York: Lippencott, 1960.

Underwood, B. J., Zimmerman, J., & Freund, J. S. Retention of frequency information with observations on recognition and recall. *Journal of Experimental Psychology*, 1971, **87**, 149-162.

Wells, J. E. Strength theory and judgments of recency and frequency. *Journal of Verbal Learning and Verbal Behavior*, 1974, **13**, 378-392.

NOTES

1. Although Underwood's frequency-attribute hypothesis has been classified as a multiple-process hypothesis by Howell (1973a) and Begg (1974), we feel that our classification under the trace-strength hypothesis is the correct one.

2. If discrimination coefficients are close to zero or one, a transformation such as Fisher's r to z transformation (see Hays, 1963) may be necessary in order that the normality assumption not be seriously violated; this was not the case with our data.

(Received for publication June 21, 1975.)