Journal of Experimental Psychology
1974, Vol. 103, No. 4, 706-716

HOW FREQUENCY AFFECTS RECENCY JUDGMENTS: A MODEL FOR RECENCY DISCRIMINATION [1]

ARTHUR J. FLEXSER AND GORDON H. BOWER [2]

Stanford University

Previous evidence that repetition of an item in a list enhances that item's recency relative to other items has been interpreted as favoring a memory strength theory of recency discrimination. However, serious doubt has been cast upon the validity of the strength theory by experiments such as that of Hintzman and Block, which instead favor a multitrace representation for repetitions of an item. The present experiments test two plausible interpretations of the effect of frequency on relative recency judgments. The first, that low frequencies result in poor recognition memory, hence poorer recency discrimination, is discounted in Experiment I, which still found sizeable frequency effects on relative recency judgments even when considering only recognized items. A multiple-trace theory of contextual time tagging was then proposed to account for the effects of event frequency on subjective recency. Experiment II, which collected event frequency as well as relative recency judgments, yielded data which were fit quantitatively by the multiple-trace time tagging theory. It was found, too, that relative "distance" judgments were not psychologically symmetric to relative recency judgments—a result not predicted by the time tagging model.

How does one judge how long ago an event took place? If the event is in the distant past, the process very likely involves relating the event to one or more temporal "mileposts" which may be assigned a date more or less directly, such as one's high school graduation. The event of unknown recency is ordered with respect to the known event by an inferential process, using the fact that if B resulted from A, then B followed A.

This explanation does not lend itself well to explaining how people can judge accurately the recency of unrelated events which occur in haphazard order over short periods that contain very little in the way of temporal landmarks (e.g., the order of arrivals for a concert or lecture). Experiments involving judgments of recency for stimulus items in a relatively homogeneous list provide examples of such a situation. Several investigators (e.g., Fozard & Yntema, 1966; Hinrichs, 1970; Morton, 1968) have proposed memory strength theories for such recency judgments. Allegedly, recency is judged on the basis of the present strength of a decaying memory trace that was established at a standard initial value at its time of formation. A corollary to this view is that multiple occurrences of the same event will increment a single cumulative strength. Thus, repeating an item in a list should make that item seem more recent than its neighbors. Fozard and Yntema (1966) and Morton (1968) have reported results showing that frequency does, in fact, have a small but significant effect on comparative recency judgments.

In both experiments, *S* judged which of two items, denoted here as A and B, he had seen more recently. If an item had appeared more than once, *S* was to consider only its most recent appearance in making this judgment. Three repetition paradigms were used, to be denoted as AB, AAB, and ABB. Here AAB, for example, is to be interpreted as Item A having two separate presentations in the list prior to the single appearance of Item B. These experiments yielded the highest percentage of correct responses (saying "B more recent") in Paradigm ABB, followed by AB, followed by AAB, as predicted by a strength

[1] This research was supported by Grant MH-13950-07 from the National Institute of Mental Health to Gordon H. Bower.

[2] Requests for reprints should be sent to Gordon H. Bower, Department of Psychology, Stanford University, Stanford, California 94305.

theory. Repetition of an item added between .05 and .10 to that item's probability of being judged the more recent.

The above results are consistent with a cumulative strength theory; however, they happen not to be discriminating because they can also be shown to be consistent with a multiple-trace hypothesis, which assumes that repetition of the same event results in separate memory traces (rather than the strengthening of a single trace). In this "multitrace" view, the trace corresponding to each separate presentation of an item will be considered to have its own "time tag" enabling estimation of the recency of that presentation. This framework does not specify the exact nature of the hypothetical time tag. One possibility, discussed by Hinrichs (1967), utilizes the idea of a trace undergoing autonomous decay, as in the strength theory, but with the stipulation that repetitions of the same event do not accumulate in a common strength measure, but rather remain as distinctly separate measures. A somewhat different view identifies a recency judgment for an event with a judgment of the duration intervening between the event and the the present. Hinrichs (1967) discusses a learning model embodying this notion, in which recency judgments are derived from a scaled-up count of remembered intervening items. A contributory factor to either of these "time-telling" mechanisms might be information retrievable from the memory trace regarding the cognitive environment or context surrounding the test item's earlier occurrence (see Anderson & Bower, 1972). For example, Hintzman, Block, and Summers (1973) have demonstrated that *S*s in multilist experiments retain information regarding whether an item was presented near the beginning or the end of a list, even when they have mistakenly assigned the item to an incorrect list. This is but one demonstration of how contextual cues might be implicated in the time-tagging process.

Whatever the mechanism, there is experimental evidence that separate recurrences of an item result in separate, noninteracting time tags. Hintzman and Block (1971) required *S*s to give absolute recency judgments for both presentations of twice-presented items. It was found that first-presentation recency judgments were unaffected by the recency of the second presentation, and vice versa. Clearly, this result is incompatible with the idea that repetition of an item increases a common strength measure on which later recency judgments are based.

In the light of this evidence, one is motivated to look for an alternative (i.e., "non-strength") explanation of the effect of frequency upon recency judgments. One possible explanation attributes the effect to recognition failures. If *S* is asked to compare the recency of two items but fails to even recognize one of them, he will surely choose the other as the more recent item. Since recognition failure is less likely for doubly presented items than for singly presented ones, this "recognition artifact" could cause doubly presented items to be judged on the average as more recent than singly presented ones. This would explain the weak effect found in the Morton (1968) and Fozard–Yntema (1966) experiments.

To test for this possible interpretation, an experiment was performed similar in design to the Morton and Fozard–Yntema studies. However, *S*s were asked to make comparative recency judgments only when they recognized both test items, thus eliminating the effects of nonrecognition. Both words and nonsense syllables were used as stimuli so as to observe these conditionalized recency judgments under conditions of both high and low recognition.

The experiment also included a BAB paradigm in addition to the AB, AAB, and ABB paradigms which had been employed in the earlier studies.

Experiment I

Method

Subjects. Thirty-seven paid *S*s, 26 females and 11 males, were recruited by newspaper advertisements from the Palo Alto area. All were between the ages of 16 and 30 yr.

Materials and apparatus. Four lists of words and four lists of nonsense consonant–vowel–consonants were prepared. The words were selected on the basis of high imagery value from a list of nouns compiled by Paivio, Yuille, and Madigan (1968), and the nonsense syllables were selected for low meaningfulness from the Krueger list reprinted by Underwood and Schulz (1960). The nonsense syllables within each list were also selected to minimize acoustic and visual similarities.

All materials were presented on slides by means of a Kodak Carousel projector. An electronic timer controlled the presentation rate.

Design. Each list was 34 items long and contained two examples of Paradigms AB, AAB, BAB, and ABB. In the AB paradigm, Items A and B were separated by 3 intervening items. In the other three paradigms, the spacings were 2 intervening items between the first and second members of the triplet and 3 intervening items between the second and third. Each list also contained 2 singly presented and 2 doubly presented items which were included for later recency comparisons with non-presented "catch" items. Three buffer items were included at the beginning and at the end of each list and were not tested.

After 19 (about half) of the *S*s had been run, the within-lists ordering of the items in each list was changed to minimize list-position effects that might favor one paradigm over another. Also, repeated words in the initial lists were made single-presentation words in the rearranged lists, and vice versa.

All items were viewed at a 2-sec. presentation rate. At the end of each list, *S*s wrote answers to 14 test questions. The questions referred to test slides consisting of two items. The test phase of the experiment was *S*-paced in that the experimenter would present a new question slide when all *S*s had finished the previous question (about 10 sec. per question).

Procedure. The *S*s were tested in groups of from 4 to 12. Each *S* was tested on all eight lists. Two different types of test form were used: In one, *S*s made comparative recency judgments; in the other, *S*s made absolute recency judgments of both test words. Each *S* used one of the two forms for his first four lists, and the other for the remaining four. Immediately prior to List 1 and to List 4, *S*s were given practice trials, consisting of 20 study items and 6 test questions, in order to acquaint them with the test form they would be using for the subsequent four lists.

On the comparative recency form, *S*s indicated whether or not each of the two items shown on the current test slide had been presented and gave a confidence rating of 1 to 3 (3 being highest) if they said that it had been presented. If and only if they had said that both items had been presented, they then chose the more recently presented item.

On the alternate type of test form, *S*s made list-position judgments on a scale of 0 to 100, with 0 denoting the beginning of the list and 100 the end. A line numbered by 10s from 0 to 100 was included at the top of this form to aid *S*s in visualizing their choices. The *S*s indicated the number of times each item had appeared (*S*s had been instructed that this could only be zero, one, or two), and gave a position judgment for each appearance. A later version of this form, used by the last 18 *S*s, required a confidence rating corresponding to each position judgment. The *S*s were instructed that this confidence rating reflected their assurance that the item had appeared on the given occasion, rather than the accuracy with which they felt they could temporally locate the appearance.

The ordering of the 14 test questions for each list followed roughly the order of the study items, thus approximating a constant interval between study and test for each item. Since test items took longer than study items, this approximation is a crude one.

Results and Discussion

Comparative recency data. The proportion of correct recency choices for recognized pairs for each paradigm is shown in Table 1. The numbers in parentheses following each proportion are the numbers of observations. Since comparative recency judgments were made only when both items were recognized, the nonsense data, which had a lower rate of recognition, have fewer observations than those for words.

The strength theory predicts the ordering ABB > BAB > AB > AAB, in terms of the proportion of correct recency judgments in each paradigm. The Table 1 data for words show a significant linear trend in this ordering, t (36) = 3.32, $p <$ .01, computed by deriving a slope for each *S*, using the arcsine-transformed proportion correct in each paradigm. The observed ordering was identical to the predicted except that the expected relationship BAB > AB failed to occur; the difference between these two paradigms was not significant, t (34) = .11, using the same method as described above.

For nonsense syllables, none of the paradigms yielded proportions differing signifi-

TABLE 1

PROPORTION OF CORRECT RESPONSES AND NUMBER OF OBSERVATIONS (IN PARENTHESES)

Paradigm	Words	Nonsense syllables
AB	.700 (110)	.580 (81)
AAB	.479 (119)	.512 (82)
BAB	.675 (120)	.593 (86)
ABB	.735 (117)	.481 (79)

cantly from the chance level of 50%. (The average standard deviation of the proportions was about .055.)

Since none of the nonsense syllable paradigms showed recency discrimination above the chance level, no significance should be accorded the ordering of these proportions. The absolute recency judgment data, to be presented later, also show that recency discrimination for nonsense syllables was practically nil in this experiment.

The word data show that frequency still influences recency even when the previously discussed influence of nonrecognition has been eliminated by requiring *S*s to recognize both test words before making a comparative recency judgment between them. Having eliminated the possible nonrecognition artifact, we conclude from Experiment I that there is a "real" frequency effect upon recency judgments.

The existence of the frequency effect presents an apparent theoretical contradiction. This effect has in the past been taken as evidence in favor of the strength theory of recency judgments. Yet powerful independent evidence, such as that provided by the Hintzman and Block (1971) experiment, exists in favor of a multitrace model, which is inconsistent with the strength theory.

This unsatisfactory state of affairs leads us to propose a different explanation for the effect of frequency on recency. We shall adopt the multitrace viewpoint, with the additional assumption that the time tagging process is subject to a certain amount of random "noise." Thus, when *S* retrieves a time tag from an event and uses it to arrive at a subjective recency estimate x, this estimate will not be correlated perfectly with the actual recency of the event. Rather, we will assume that for a fixed actual recency, x will have a probability distribution about some central value $\bar{x}$. Therefore, two events which are relatively close together in time, so that their x distributions overlap, will not always be perceived as having occurred in the correct order, even if *S* has retrieved a time tag (of imperfect accuracy) for both. Let x_1 and x_2 denote the perceived recency values for two events, A and B, where B follows A as in the AB paradigm of Experiment I. (We assume for the moment that *S* has succeeded in retrieving a time tag for both events—a situation which is not always the case.) Also, let $f_1(x)$ and $f_2(x)$ denote the probability density functions associated with these two judgments. If greater values of x are associated with more recent events, then we may write:

$$p(\text{B judged more recent than A}) = p(x_2 > x_1) = \int_{-\infty}^{+\infty} f_2(x) F_1(x) dx, \qquad [1]$$

where $F_1(x)$ denotes the cumulative distribution function of x_1.

When there are three x distributions involved, as is the case with the triplet paradigms AAB, BAB, and ABB, Equation 1 may be generalized as:

$$p(\text{item from distribution } i \text{ judged most recent}) = p(x_i > x_j \text{ and } x_i > x_k) = \int_{-\infty}^{+\infty} f_i(x) F_j(x) F_k(x) dx, \qquad [2]$$

where we arbitrarily denote the three subjective recencies as x_i, x_j, and x_k. If the three x distributions were known, this equation could be used to predict the proportion of correct responses in the three triplet paradigms. For example, the probability of *incorrectly* choosing Item A as most recent in the BAB paradigm is equivalent to the probability of choosing the item from Distribution 2 as most recent.

It may also be shown that the model predicts the ordering ABB > BAB > AB > AAB in terms of p(correct) in each paradigm.[3] This ordering is identical to that predicted by the strength theory and is also consistent with the results for words in Table 1. In intuitive terms, the model holds that the frequency effect is a consequence of the fact that in Paradigm AAB,

[3] This ordering rests on the assumption that discriminability between Distributions 1 and 2 is comparable to that between 2 and 3, and that A and B in the AB paradigms are of recencies corresponding to Positions 2 and 3 in the triplet paradigms.

S must successfully distinguish that B was more recent than both As, whereas in Paradigm ABB it is sufficient that *S* realize that *either* B followed the A.

The above ordering was derived under the assumption that a time tag was retrieved for each presentation of Item A and of Item B. Although *S* was required to recognize A and B before making a recency judgment in Experiment I, this does not ensure that a time tag was always retrieved. Even given a recognition response, this retrieval assumption could still be wrong in two types of circumstances.

First, *S* would be able to claim recognition even if he retrieved only one tag for a doubly presented item. The model predicts that cases of this type will result in a general weakening of the predicted frequency effect (i.e., to the extent that only one tag of a doubly presented item is retrieved, the condition theoretically becomes like AB). In Experiment II, to be described, frequency estimates for Items A and B were taken in order to help us identify test situations of this type, where we adopt the viewpoint that a frequency estimate reflects the number of tags retrieved for that item.

The second case that violates the time-tag retrieval assumptions results from false recognitions. In Experiment I the false recognition rate was .15 for words and .47 for nonsense syllables. We attribute such false alarms to the faulty retrieval of a random but inappropriate context tag due to confusion of the test stimulus with similar stimuli presented earlier. Presumably, *S* arrives at a "perceived recency" estimate for such falsely recognized items just as he does for an item with a correctly retrieved tag. This process will affect not only the new (catch) items but can also contaminate recency judgments for presented items, resulting in some intrusions of recency values drawn from inappropriate distributions.

Both of these cases in which the retrieval assumption is not valid would be expected to occur much more frequently for nonsense syllables than for words. Therefore, in retrospect, it is understandable that the nonsense syllable data did not show a consistent effect of frequency on recency.

The applicability of the model to the data of Experiment I is also vitiated by the fact that the study–test interval was constant only to within a very rough approximation. Experiment II was designed to remedy this defect as well as to minimize the problems caused by failures of the retrieval assumption. Before describing this experiment, however, we will present some brief results from the absolute recency judgment task of Experiment I.

Absolute recency data. Figure 1 shows the relationship between recency judgments and list position for words which were presented once and for which a single recency judgment was given. These judgments were made on a scale of 0 to 100, with 0 denoting the beginning and 100 the end of the study list. Recency discrimination, as measured by the slope of the regression line, was significantly better for words than for nonsense syllables, *t* (1,496) = 5.90, $p < .001$. The mean recency judgment for words (52.0 on a scale of 0–100) was slightly but significantly greater than that (48.5) for nonsense syllables, *t* (1,498) = 2.56, $p = .01$. Although the regression coefficient for the nonsense syllables was significantly different from zero, *t* (656) = 5.72, $p < .001$, recency discrimination for nonsense syllables was very poor in this experiment. One demonstration of this is that the increase in mean judged recency between the first and last positions tested was only about 7 points—from 44 (Position 2) to 51 (Position 33).

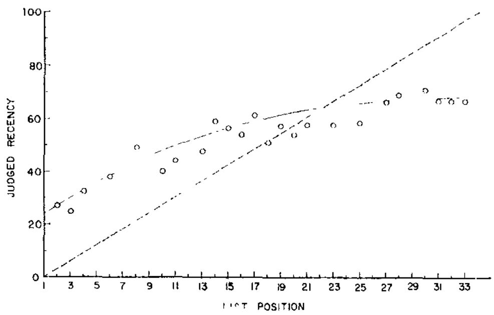

FIGURE 1. Mean judged recency as a function of list position for words. (Dotted line corresponds to correct judgments.)

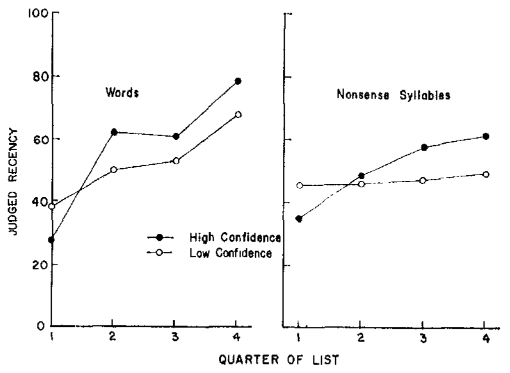

FIGURE 2. Mean judged recency for high- and low-confidence words and nonsense syllables.

We propose that most of this difference in recency discriminability between words and nonsense syllables is attributable to differing amounts of contamination by false alarms which arise from generalization errors—that is, by confused retrieval of a tag other than the correct one for the test item. This view implies that recognition performance and recency discrimination for different types of stimulus items should be closely associated. Data reported by Fozard and Weinert (1972) support this view. In their experiment recency judgments for pictures and for nouns were compared, with results closely resembling the relationship found in our experiment between words and nonsense syllables. That is, the stimulus material (namely, pictures) that had the better recognition rate also showed the stronger dependence of judged recency upon actual recency.

Similar considerations lead to the prediction that recency discrimination will be superior for items recognized with high confidence since false alarm contamination should be low in such cases. Figure 2 shows recency judgments separately for items assigned high- and low-confidence ratings in Experiment I[4]. These are plots for items which were presented once and for which a single recency judgment was made. The list positions have been categorized into first through fourth quarters for these graphs.

The slope for the high-confidence items was significantly greater than that for the low-confidence items, both in the case of words, t (240) = 2.49, $p < .02$, and in the case of nonsense syllables, t (192) = 2.22, $p < .05$. Mean recency judgments in all categories were close to 50 (on a scale of 0–100), with no significant differences observed.

The data presented in Figure 2, while consistent with our model, serve to rule out the plausible a priori hypothesis that recency judgments and confidence ratings are derived from a common underlying dimension. These data show that stimuli recognized with a high degree of confidence are not judged to be more recent than their low-confidence counterparts; rather, it appears that high confidence is associated with greater accuracy of the recency judgment, whether it be long or short.

Catch items that were erroneously given a frequency judgment of one were assigned mean recency judgments of 39.2 and 48.2 for words and nonsense syllables, respectively. The difference between these two means is significant, t (185) = 2.04, $p < .05$. This difference has several interpretations, but the data are insufficient to dwell upon the matter.

EXPERIMENT II

Experiment II was designed to provide a quantitative test of the model's success in predicting comparative recency judgments for words under steady-state conditions. This experiment obtained frequency judgments for each pair of test items as well as comparative recency judgments, so that data from the recency task could be conditionalized upon frequency judgments. (In Experiment I, comparative recency judgments had been accompanied only by a judgment of whether the items were new or old.)

A second aim of Experiment II was to investigate performance in a comparative

[4] The high- and low-confidence items shown in Figure 2 are those which were assigned confidence ratings of 3 and 1, respectively. The results from the 2-rated items are omitted for clarity. They lie generally between the 1- and 3-rated items, and are intermediate in slope.

distance task, in which *S*s were asked to choose the more distant item rather than the more recent. The reasoning which predicted the ordering ABB > BAB >AB > AAB in terms of proportion of correct responses in a comparative recency judgment task predicts just the reverse ordering of conditions *if* the task of choosing the more distant item is just the converse of choosing the more recent item.

Method

Subjects. Forty-three *S*s, 24 females and 19 males, ranging in age from 18 to 31 yr., participated. The *S*s were either paid for their participation or received course credit.

Materials and apparatus. Three lists of words were prepared, using nouns of high imagery value selected from the list compiled by Paivio et al. (1968).

All materials were presented on slides by means of a Kodak Carousel projector controlled by an electronic timer.

Design. Each list was 240 slides in length and contained seven examples of the AAB, BAB, and ABB paradigms, nine examples of singly presented words which were tested against catch (nonpresented) items, and six examples of doubly presented words which were tested against catch items. In addition, two different AB paradigms, differing in their interitem spacings, were used. One spacing corresponded to the long spacing between Item 1 and Item 3 in the triplet paradigms (AAB, BAB, and ABB); the other corresponded to the short spacing between Item 2 and Item 3, which was made equal to the spacing between Item 1 and Item 2. Each list contained eight examples of the AB_S paradigm (i.e., Paradigm AB using the shorter of the two intervals), and nine examples of the AB_L paradigm (the longer interval). Each list also contained 17 filler items which were never tested; these included 6 buffer items placed at the beginning of the list. The 1–2 and 2–3 intervals in the triplet paradigms both contained 5 intervening items, and the study-test interval in all cases was 20 items, as measured from the last item in the paradigm. The numbers of items intervening between A and B in the AB_S and AB_L paradigms were 5 and 11, respectively. All of these intervals were allowed to very within ±1 of the specified lengths in order to simplify construction of the lists. In measuring these intervals, test questions, consisting of a test slide followed by a blank slide, were counted as 2 items.

All three lists had the same structure, except that the positions of the AAB, BAB, and ABB paradigms were rotated from list to list. The order in which the three lists were presented was counterbalanced across *S*s.

Procedure. The *S*s were tested in groups of between 2 and 12. Each *S* was tested on all three lists. Data from one list had to be discarded for 6 *S*s due to equipment failure during testing.

The *S*s first read instructions for the experiment which specified that no item would appear more than twice, that repetitions would always be within a dozen items of one another, and that *S*s would never be tested on any item older than a lag of 30. They were then tested on a practice list identical in structure to the experimental lists, but only one slide tray in length rather than three. The rate of presentation was 4 sec. per slide. This rate allowed 8 sec. for answering the test questions, since each test slide was followed by a blank slide.

Half of the *S*s were instructed to choose the more recent item during the practice test, while the other half chose the more distant item. Thereafter, *S*s alternated between the two tasks over successive lists.

The *S*s wrote their answers; each question called for a frequency judgment for each of the two items presented on a test slide and a comparative recency (or comparative distance) judgment between the two items. The *S*s made the recency or distance judgment only if they had assigned a nonzero frequency to both items.

Results and Discussion

Recency judgments. The proportions of correct recency judgments in the various paradigms are shown in the first column (C_1) of the left half of Table 2. As predicted by the model, these proportions are ordered AAB < AB < BAB < ABB, showing a clear effect of event frequency upon recency judgments.

Since a goal of Experiment II is to provide a quantitative test of the model, we must now deal explicitly within the model with the problem of retrieval failures, i.e., the idea that *S* may or may not retrieve a time tag corresponding to each presentation of each item. We will argue below that the effect of such retrieval failures can be greatly minimized by the simple procedure of examining only those comparative recency judgments made after *S* first judged the correct frequency for both items being compared.

In the foregoing, we have been considering a time tag as one attribute of the distinct memory trace left by each presentation of an item. We will assume in what follows that *S*'s frequency estimate for an item provides a count of the *number* of memory traces (and their associated time tags) he has retrieved for that item. (For

TABLE 2

PROPORTION OF CORRECT RESPONSES IN EIGHT FREQUENCY PARADIGMS

Paradigm	Recency judgments					Distance judgments				
	C_1	No. of observations	C_2	No. of observations	P	C_1	No. of observations	C_2	No. of observations	P
AB_S	.588	376	.610	310	.614	.526	380	.546	315	.614
AB_L	.623	400	.641	345	.654	.580	417	.588	352	.654
AAB	.568	366	.546	229	.477	.602	374	.669	242	.791
BAB	.611	368	.689	225	.712	.530	336	.564	204	.641
ABB	.675	366	.766	239	.765	.517	360	.516	225	.431
AAB_1			.622	90	.634			.448	87	.634
BAB_1			.492	124	.536			.495	101	.464
ABB_1			.516	97	.598			.524	105	.598

Note. Abbreviations: C_1 = conditionalized upon *S*s recognizing both A and B; C_2 = conditionalized upon *S*s giving correct frequency judgments for A and B in the first five paradigms, or giving both A and B frequency judgments of one in the last three; P = minimum χ^2 predictions for C_2 generated by model, using parameters estimated from recency judgments only.

now, this assumption must rest on its intuitive appeal; a forthcoming paper concerning frequency discrimination will present evidence favoring this interpretation.) The "perfect retrieval" assumption of the model is contaminated by two types of failure, as discussed earlier. In the first type, *S* fails to retrieve a memory trace associated with the test stimulus, while in the second type he retrieves an irrelevant memory trace. We will refer to these two types as misses and false alarms, respectively. If *S* makes a correct frequency judgment for an item, the only instances in which the retrieval assumption will be incorrect are those in which the number of misses equals the number of false alarms to that item. We now show that such fortuitous circumstances are very unlikely in the data of Experiment II.

In Experiment II *S*s made correct frequency estimates for 74% of all items. This level of accuracy suggests that a reasonable maximum value for the probability of missing a trace would be about 25%; a similar maximum may be inferred for the probability of generating a false alarm. Therefore, the probability of a correct frequency estimate due to one miss counterbalancing one false alarm would be at most $(.25)^2$, or about .06. The probability of more than one miss counterbalancing an equal number of false alarms is even more negligible.

Thus, if we use only those data from Experiment II in which *S* made correct frequency judgments for both items being compared, we can expect the retrieval assumption to be satisfied in all but a small fraction of cases. The proportions of correct recency judgments in each paradigm, subject to this conditionalization, constitute the first five entries in Column C_2 of Table 2. The conditionalization upon correct frequency judgments more than doubles the apparent magnitude of the frequency effect, as measured by the difference in proportion of correct recency judgments between the best and worst paradigms. It is these conditionalized data which we will use to test the quantitative predictions of the model, neglecting the effect of the relatively rare failures of the retrieval assumption. In addition, we will also use the model to predict the proportion of correct responses in the triplet paradigms for those cases in which *S* has erroneously assigned frequency judgments of one to both items. We will use the notation AAB_1, BAB_1, and ABB_1 to denote these cases; the proportions of correct recency judgments in these cases constitute the last three entries in Column C_2 of Table 2.

Quantitative test. To generate predictions we assume that the *x* scale of subjective recencies can be set up in such a way as to make all distributions normal with equal variance.[5] We may also normalize the scale so that variances are equal to unity, and choose the zero point at $\bar{x}_1$, the mean of the distribution corresponding to the earliest (farthest back in time) of the three recencies used. On this scale, we denote the means of the x_2 and x_3 distributions as z_1 and z_2, respectively. These two parameters are free variables which are estimated in order to fit the C_2 proportions of Table 2.

In calculating the probability of a correct response in paradigms AAB_1, BAB_1, and ABB_1, in which A and B have both been assigned frequencies of one, we have assumed that *S* is equally likely to have forgotten either of the two presentations of the doubly presented item. The missing trace leaves *S* with the functional equivalent of either AB or BA. Therefore, the probability of success in each of these paradigms may be obtained by averaging the probabilities of choosing B as the more recent in each of the two equally probable sequences that can result from missing one trace of the repeated item.

Values of the parameters z_1 and z_2 were calculated iteratively to provide a minimum χ^2 fit to the C_2 recency data. The predicted values of p(correct) for each paradigm are given in Column P on the left side of Table 2. The best fit was obtained with $z_1 = .150$ and $z_2 = .559$, yielding an insignificant $\chi^2(6) = 9.03$, $p > .15$.

These best-fitting parameter values are somewhat questionable, however, since they suggest that the discriminability between Positions 2 and 3 ($z_2 - z_1$) is much superior to that between Positions 1 and 2 (z_1), despite the fact that Positions 1, 2, and 3 were equally spaced lags in the steady-state task. Fortunately, this is not a serious defect of the model, since a satisfactory fit to the data can be achieved with z values that conform more closely to prior expectations. For example, if we require that $z_2 - z_1 = z_1$ (equal spacing), the value $z_1 = .295$ yields χ^2 (7) = 13.5, which is not significant at the .05 level. So a range of reasonable pairs of z_1, z_2 values will fit the observed proportions of correct responses in the recency judgment task fairly well.

Distance judgments. The data from the distance judgment task are shown in the right half of Table 2. Qualitatively, these data are in accord with the model: The paradigms are ordered in the reverse fashion from the recency data, as expected. Also, conditionalizing upon correct frequency estimates (Column C_2) enhances the size of the frequency effect, as was the case with the recency judgments.

The model clearly does not give an adequate quantitative account of these distance judgments. The values of z_1 and z_2 estimated from the recency data provide a very poor fit to the distance data (see Column P). Moreover, no other values of these parameters can provide an acceptable fit. The reason for this failing lies in the fact that distance judgments do not seem to be the exact psychological converse of recency judgments, as we had assumed. This fact is made apparent by a comparison of the C_1 columns for the recency and distance tasks. This comparison reveals that distance discrimination was generally poorer than recency discrimination, as shown by the fact that the proportion of correct responses in the distance paradigms averaged about .06 less than in the recency paradigms. This overall difference is significant, t (42) = 3.40, $p < .01$. The decrement was evident even in the AB paradigms, t (42) = 2.18, $p < .05$, which is quite surprising in view of the fact that the only difference between making a recency versus a distance judgment in these cases is that the opposite answer is required. Note that adding .06 to the values in the C_1 (distance) column brings all pro-

[5] It might be argued that the equal variance assumption is inappropriate, since there is evidence that absolute recency judgments increase not only in magnitude but also in variability as the judged interval becomes larger (Hinrichs, 1967). In the present experiment, where the study–test interval was appreciably greater than the interitem intervals within each paradigm, we expect that the variance of subjective recency estimates will not depend heavily on whether an item appeared in Position 1, 2, or 3. To the extent that these variances are position-dependent, the model will compensate by compressing the x-scale at the longer delays.

portions quite close to the values that would be obtained using the C_1 (recency) column and symmetry considerations. For example, if we add .06 to the proportion correct in the ABB distance paradigm (.517), we obtain a figure close to that which was observed in the AAB recency paradigm (.568).

We are puzzled by the fact that comparative distance judgments appear to be more difficult than comparative recency judgments. Of the 43 *S*s, 29 did better on the recency task—a proportion significantly higher than half, $z = 2.14$, $p < .05$.

The asymmetry of the two judgmental tasks is reminiscent of that found by Audley and Wallis (1964), who asked their *S*s to judge which of two illuminated areas was either "lighter" or "darker." When the areas were lighter than the background, the question "Which is lighter?" elicited faster responses than the question "Which is darker?"; however, just the opposite results were obtained when the areas were darker than the background. Analogous results were also obtained using pitch judgments (Wallis & Audley, 1964). If *S*s in our experiment considered the stimulus items as relatively recent in comparison to some "background" of events (e.g., those prior to the experiment), the Audley–Wallis conclusion would imply that recency judgments would be easier than distance judgments, as was observed. Although the Audley–Wallis studies dealt with latencies rather than with proportions of correct responses, it is nevertheless tempting to draw a parallel between the two. For example, if comparative distance judgments involve an extra stage of processing for some *S*s, one might expect to find both longer latencies and higher error rates.

General Discussion

Let us summarize our arguments, results, and conclusions. Previous results, that an event's frequency enhanced its subjective recency, suggested a trace strength model of recency judgments. Allied against this strength model are several results (e.g., Anderson & Bower, 1972; Hintzman & Block, 1971) indicating that *S*s can keep separate track of the several contexts in which a given item has occurred; such results strongly suggest a "multiple-trace" hypothesis regarding the effects of repetition of an event.

How then are we to explain in these terms the effect of an event's frequency upon its apparent recency? Experiment I investigated the possibility that the previously found frequency effect was a trivial consequence of the known effect of event frequency upon recognition memory. However, the results of Experiment I, in which recency judgments were conditionalized upon recognition, failed to support this conjecture; even considering only recognized items, more frequent events were still apparently more recent. But, as before, several aspects were incommensurate with the trace strength model, e.g., greater confidence of recency judgments was reflected in the absolute accuracy rather than the recency per se of the judgments.

An alternative model of recency judgments was formulated, utilizing the notion that each repetition of an event produces a separate memory trace containing elements summarizing the prevailing context at the time of encoding (see Anderson & Bower, 1972). Among the attributes of such a context marker would be information allowing an imprecise estimate of its recency—a "time tag." The presumed "noisiness" of such recency judgments was represented by a mathematical formulation along the lines of Thurstone's (1927) theory of discriminal dispersions. Such a model implies that judged frequency will be a critical determinant of relative recency judgments. The results of Experiment II confirmed this prediction. The effects on recency of variation in event frequency were largest when event frequencies were correctly estimated; when a doubly presented item was later judged to have been only once presented, it was chosen as more recent with a proportion approaching those of the appropriate once-presented items. These and related effects were predicted quantitatively from the multiple-trace model. It was noted, too, that the question" Which event is more distant?" was more difficult than its logical converse, "Which event is more recent?"—a result reminiscent of asymmetries in other judgmental tasks.

The main conclusion from Experiment II is that the effects of frequency on recency judgments, as well as on recency judgments conditionalized upon frequency estimates, are all predictable (quantitatively) from the multiple-trace time-tag model. Such a model is also required to fit other results on frequency estimates. Our overall intent is thus to promote the multiple-trace time-tag model as adequate to deal with both subjective frequency and recency judgments based on memory for events.

REFERENCES

Anderson, J. R., & Bower, G. H. Recognition and retrieval processes in free recall. *Psychological Review*, 1972, **79**, 97–123.

Audley, R. J., & Wallis, C. P. Response instructions and the speed of relative judgments: I. Some experiments on brightness discrimination. *British Journal of Psychology*, 1964, **55**, 59–73.

Fozard, J. L., & Weinert, J. R. Absolute judgments of recency for pictures and nouns after various numbers of intervening items. *Journal of Experimental Psychology*, 1972, **95**, 472–474.

Fozard, J. L., & Yntema, D. B. The effect of repetition on the apparent recency of pictures. *American Psychologist*, 1966, **21**, 879. (Abstract)

Hinrichs, J. V. Judgment of recency: Data and theory. Unpublished doctoral dissertation, Stanford University, 1967.

Hinrichs, J. V. A two-process memory strength theory for judgment of recency. *Psychological Review*, 1970, **77**, 223–233.

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

Hintzman, D. L., Block, R. A., & Summers, J. J. Contextual associations and memory for serial position. *Journal of Experimental Psychology*, 1973, **97**, 220–229.

Morton, J. Repeated items and decay in memory. *Psychonomic Science*, 1968, **10**, 219–220.

Paivio, A., Yuille, J. C., & Madigan, S. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology Monograph*, 1968, **76**(1, Pt. 2).

Thurstone, L. L. A law of comparative judgment. *Psychological Review*, 1927, **34**, 273–286.

Underwood, B. J., & Schulz, R. W. *Meaningfulness and verbal learning*. Philadelphia: Lippincott, 1960.

Wallis, C. P., & Audley, R. J. Response instructions and speed of relative judgments: II. Pitch discrimination. *British Journal of Psychology*, 1964, **55**, 133–142.

(Received February 9, 1974)