Interactions of Stimulus Attributes, Base Rates, and Feedback in Recognition

W. K. Estes and W. Todd Maddox
Harvard University

Continuous old–new recognition was studied in relation to 3 factors that have been relatively neglected in previous research—stimulus attributes, old–new base rates, and informative feedback following responses. Under all conditions, both hits and false alarms increased over trials and all measures of recognition depended strongly on stimulus properties, notably interitem similarity. In contrast to expectations based on earlier results, both hit and false-alarm levels proved independent of old–new base rate when tests were given without feedback; with feedback added, false-alarm rates tended to approach true old-stimulus base rates with some types of stimuli, though not with words. The findings are compatible, in general, with current composite-memory models and were predicted in detail by an array-similarity model deriving from categorization theory.

Recognition is currently a focal area for comparing and testing models of memory. A cluster of recent studies (Hintzman, 1988; Murdock, 1991; Shiffrin, Huber, & Marinelli, 1995; Shiffrin, Ratcliff, & Clark, 1990; Yonelinas, Hockley, & Murdock, 1992) has been fruitful in two respects. First, it has added substantial support for a class of global models in which recognition of an item depends on the total activation it generates in a composite memory rather than on its degree of match or mismatch with individual representations retrieved by a search process. Second, this research has clarified empirical issues concerning the interrelated effects on recognition of several commonly studied variables, in particular, the memory-trace strength of a test item and the length and composition of the list in which the item was studied.

However, the scope of these studies is limited in some ways that deserve attention. One is the range of stimulus materials. All of the cited studies, as well as other substantial bodies of research bearing on recognition models (e.g., those reviewed by Glanzer, Adams, Iverson, & Kim, 1993), almost exclusively use common English words as stimuli. The convenience of using words as stimuli and the standardization this practice affords come at a cost, for it is not clear how well the results obtained and the models supported will generalize to the many other kinds of stimuli that are used in memory research and that are important in life outside the laboratory. We are not implying that all varieties of stimuli need be represented in recognition research, but it is important that stimulus proper-

ties known to be relevant to recognition be examined over adequate ranges. Two such properties are interitem similarity and item reheasability, and common words are at the high extreme on the latter and near the low extreme on the former. In this study, we compare recognition of words with recognition of two kinds of stimuli commonly used in memory studies that differ greatly from words with respect to both similarity and reheasability.

A second important limitation has to do with experimental design. From the turn of the century, two designs have predominated in recognition research, study–test (exemplified in an early study by Strong, 1912) and continuous recognition (introduced by Shepard & Tegotsoannyen, 1961). In the study–test method, which mimics situations common in school learning and testing, a person first studies a list of items, then is given a test for recognition of the studied items. In the continuous recognition method, which seems to typify many experiences outside the laboratory when one is learning to recognize people, places, objects, a person views and responds "old" or "new" to each member of a long sequence of items; each item after the first may be either old (having occurred earlier in the sequence) or new, usually with fixed probabilities (base rates).

The study–test design is more convenient for most experimental purposes, and in its pure form is easier to interpret in terms of models. However, the pure form is rarely seen except in the familiar paradigm introduced by Sternberg (1966) in which a person views a short list of items (most often random digits), then is tested with a single item from the list or a single new item. Much more common, because it yields much more data on recognition of a given list, is a hybrid design in which a person views a study list that may contain up to several hundred items, then is tested by means of a continuous recognition sequence in which old items (all drawn from the study list) are interspersed with distractors (new items). The hybrid design, used in all of the studies cited above, adds a complication in that learning, that is, the storage of representations of perceived items in memory may continue from the study list through the test series. For simplicity, in model-
oriented research this complication is usually treated simply as a nuisance, and average scores from responses to old items and responses to new items in the test series are the basis for the measures of recognition that are to be predicted by the models. More rigorous tests of models could be achieved if the task set for them were also to account in detail for the way in which measures of recognition vary over the test series.

In this study, we undertake this more exacting task with a modification of the standard hybrid design that permits us to trace over the test series measures of recognition both for items whose lag from study to test increases over the series (the usual procedure) and for items whose lag is constant over the series. Also, we manipulate two factors that must be expected to influence performance in continuous recognition but about which surprisingly little is known—old–new stimulus base rates and feedback conditions.

Most often, in both continuous recognition studies and in the test sequences of studies using hybrid designs, old and new items occur equally frequently. However, other relative frequencies (base rates) are occasionally used. From time to time following the ground-breaking study of Shepard and Teghtsoonian (1961), there have been suggestions that probabilities of false alarms (giving "old" responses to new stimuli) tend to approach the true old-stimulus base rate, a particularly striking instance appearing in a study by Olson (1969). In the framework of models based on signal detection theory, this apparent tendency has been interpreted to be a consequence of peoples' adjustments of their criteria for giving "old" responses (Donaldson & Murdock, 1968; Murdock, 1974; Parks, 1966). We say "apparent tendency" because only case we have found in the recognition literature in which base rate has been varied within an experiment is a recent study by Ratcliff, Shen, and Gronlund (1992), which was conducted with a study–test procedure and only one type of stimulus material. Thus, it appears that actually very little is known about the relation between false-alarm levels and stimulus base rate.

With respect to feedback, the standard procedure in studies of recognition has been to present stimuli on both study and test trials without feedback following responses. In a few studies, for example, Murdock and Lamon (1988), feedback has been given in the form of information about correctness of responses, but we have found no case in which feedback conditions have been varied within an experiment.

In this study, we redress matters by varying both old–new stimulus base rate and feedback conditions within experiments and by examining the effects of these variables with several types of stimuli. Motivation for the study arose in part from a desire to test an initially surprising prediction from a class of recognition models deriving from categorization theory. As we show in a later discussion of these models, they imply that the tendency suggested in the earlier literature for false-alarm rate to match old-stimulus base rate cannot hold in general and that, in the absence of feedback (or relevant prior knowledge), false-alarm rate should be independent of base rate.

More broadly, this study is one of a series planned to further the development of models that treat recognition as a special case of categorization. One class of these models originated with the observation that instance-based models of category learning can be reinterpreted, almost without modification, to apply to recognition (Estes, 1968b; Medin & Schaffer, 1978; Nosofsky, 1988). Another closely related class derives from adaptive network models of categorization (Gluck & Bower, 1986; Nosofsky, Kruschke, & McKinley, 1992). In work immediately preceding this study, simple versions of the two models were formulated and applied to some of the currently popular experimental paradigms bearing on effects of study–list length, list composition, and prior learning experiences (Estes, 1994). Henceforth, for brevity, we refer to these models as the array model and the similarity network model. Our purpose now is not to compare the fits of these categorization-based models with others in applications to the familiar paradigms, but, rather, to take advantage of some properties of these models that make them especially appropriate to aid us in investigating how recognition depends on the experimental factors that are studied in the experiments to be described.

Overview of the Study

The study comprises two experiments designed to enable examination of old–new recognition as a function of feedback condition, stimulus type, and old–new stimulus base rate. To make it possible to maintain constant base rates, each recognition test series was preceded by a study series in which experimental participants passively viewed a sequence of all-different stimuli. We chose the stimuli both to achieve a wide range of similarities and to include the types most often used in preceding research: For the type expected to have the highest level of interitem similarity, we used random digit triads (as in Shepard & Teghtsoonian, 1961); for the intermediate level, we used random consonant triagrams (as in Olson, 1969); and for the lowest level, we used common English words, the preponderant choice in recent work on recognition (Glanzer & Adams, 1990; Murdock & Lamon, 1988; Ratcliff, Clark, & Shiffrin, 1990; Yonelinas et al., 1992). We have found no reports of research on psychological similarities of these types of stimuli, and we do not assume that the types differ solely with respect to similarity. Rather, we use the array model to help determine how the stimulus types differ on both similarity and other attributes relevant to recognition.

We originally planned an experiment (Experiment 1) in which three feedback procedures and two base rates were

---

1. Over a 500-trial series with consonant triagrams as stimuli, there were approximately 40% old stimuli in every 50-trial block after the first, and the observed false-alarm rate rose from approximately 20% in the first block to 40% over the last 5 blocks.
2. When planning this study, we intended to make a detailed comparison of predictions from the array and similarity network models, and we did, in fact, carry out complete, parallel analyses of the data of the study in terms of the two models. We do not report those for the network model here because the fits of the two models to our data proved to be almost identical. On reflection, we realize that this result might have been anticipated because in our experiments items were never repeated more than once, and therefore the differences between the learning mechanisms did not come into play. For situations in which larger numbers of repetitions or variations in study time per item occur, these models yield testably different predictions (Estes, 1994, Ch. 7).
combined factorially with letter versus digit stimuli. Once started, we recognized that the range of similarity was not great enough for our purposes, so we added word stimuli, again combined factorially with the other two factors. Preliminary analysis of the data of this experiment left us uncertain as to whether the recognition test series had been long enough to allow evaluation of asymptotic false-alarm rates, so we conducted a complete replication (Experiment 2) with the length of the test series doubled. In this presentation, we describe the methods of the individual experiments in the following section. We then organize the results in terms of effects of the principal factors rather than in terms of experiments.

General Method

Experiment 1

Participants. The participants were 48 Harvard undergraduates who were paid for their service.

Apparatus. Stimulus displays and informative feedback were presented on the screen of a Macintosh I11 microcomputer. Participants' responses were entered on the keyboard, and data were recorded automatically.

Design. In what we term the digit-letter condition, the principal experimental variables were base rate (relative frequencies of old and new stimuli) and feedback procedure. The base rates were 33% and 67% old stimuli during the continuous recognition series. Feedback conditions were termed FB, representing informative feedback on each trial plus a bonus for high performance; NFB, representing no feedback but the same bonus; and NFNFB, representing no feedback and no bonus. These variables were crossed factorially, with 18 participants assigned randomly, 3 to each cell. Each participant took part in two replications of the experiment, the first with digit strings and the second with consonant strings as stimuli. The entire design was replicated with 18 more participants, the replications differing only in the order of occurrence of stimuli within blocks.

The continuous recognition series consisted of 192 trials. To maintain the desired old/new base rates throughout the recognition series, we modified the study—test paradigm used by Murdock and Anderson (1975). Participants were presented at the start of each replication with a list of 72 stimuli termed the study list. These stimuli were of the same type as those of the continuous recognition test that followed and were presented singly just as those of the test series. Participants were instructed to make a "new" response to each item, and all participants were treated alike in receiving no feedback and no bonus for responses to the study items. These stimuli were then available to be used as old stimuli from the start of the recognition series. The 192 trials of the recognition series were partitioned into eight 24-trial blocks. New stimuli, study-list stimuli, and repetitions of items from the preceding test block occurred with the same frequencies in each 24-trial block of the recognition series—16, 4, and 4, respectively, for the 33% base rate and 8, 8, and 8 for the 67% base rate (except for the first block, in which all old stimuli necessarily came from the study list).

The digit stimuli were random trials, drawn without replacement from the set of all possible trials in the range 111-999; the consonant stimuli were random trials drawn from the set of all English consonants, excluding y. The digit and consonant stimuli were yoked so that a given digit trial and its "mate" in the set of consonant triads occurred on the same trials.

In what we term the word condition, 12 participants were tested in a similar design but with common English words as stimuli. Because the digit-letter condition revealed no significant differences between the NFB and NFNFB data, the word condition was limited to FB and NFNFB (henceforth denoted control') procedures, which were crossed factorially with the 33% and 67% base rates, 3 participants being assigned to each cell. Each participant served in two replications with different sets of words. The word stimuli were sampled randomly from a pool of 1,000 words obtained from the Toronto Word Pool (courtesy of Michael Kahana and Bennet Murdock).

Procedure. For the study list, participants were instructed to respond "new" to each item and try to remember the items for a later test. For the recognition series, instructions were to respond "old" to each item judged to have been previously seen in either the study list or the recognition series, and to respond "new" otherwise.

During both the study and the recognition test series, the stimulus display was terminated by the participant's pressing of the Old or New key on the console; then, for the FB condition only, feedback information appeared for 1 s in the form of a capital O or N, together with a numeral denoting the current difference between correct and incorrect response frequencies. There was a 1-s blank interval between the termination of the stimulus display in the study series or end of the feedback display in the recognition test series and the onset of the next trial.

Participants in both the FB and NFB conditions were informed that they would receive a bonus payment at the end of the experiment, the amount depending on the difference between total correct and total incorrect responses over the recognition series. The amount received was typically in the range of $2.00-$4.00, added to the base payment of $5.00 received by the control participants.

Experiment 2

The method was the same as that of Experiment 1 except for the following modifications. The principal change was that the recognition test series was extended from eight to sixteen 24-trial blocks. As a consequence, to allow maintenance of constant old stimulus base rates over blocks, the study list was increased from 72 to 136 items. Also, because it was impractical to have each participant tested in two replications with the lengthened series, stimulus type in the digit-letter condition, digit or consonant triads, was changed from a within-subject to a between-subjects factor. Furthermore, because the NFB and NFNFB procedures in Experiment 1 yielded no significant differences in any aspect of the data, the NFB procedure was omitted from Experiment 2; the NFNFB procedure is hereafter denoted control. The four factors—stimulus type (digits or consonants), base rate (33% or 67%), feedback (FB or control), and order (two randomizations of the item sequence)—were combined factorially with 3 new participants assigned to each cell, resulting in a total of 48 participants.

In a word condition conducted with 12 additional participants, the method was identical to that of the word condition of Experiment 1 except that, because of the length of the session, participants were tested in only one replication, with a single stimulus order.

Because the order of occurrence of particular stimuli is likely to influence performance, controlling for order effects presents a design problem. One solution is to assign a different order to each participant; however, this procedure has the drawback that order effects are confounded with error in statistical analyses with consequent loss of power. The only feasible alternative, and the one we have used, is to assign different orders to subgroups of participants and treat order as a factor in statistical analyses. The decision concerning number of subgroups involves various trade-offs, and our purposes were best served by choosing two subgroups so that we would be able to reveal differences in the abilities of alternative models to handle specific order effects.
Table 1
Percentages of "Old" Responses to Each Stimulus Type, Averaged Over Order and Blocks (Experiment 1)

<table>
<thead>
<tr>
<th>Source</th>
<th>33% Base rate</th>
<th>67% Base rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Feedback</td>
<td>Control</td>
</tr>
<tr>
<td>Digits</td>
<td></td>
<td></td>
</tr>
<tr>
<td>New</td>
<td>30</td>
<td>42</td>
</tr>
<tr>
<td>Study</td>
<td>46</td>
<td>54</td>
</tr>
<tr>
<td>Series</td>
<td>73</td>
<td>74</td>
</tr>
<tr>
<td>Letters</td>
<td></td>
<td></td>
</tr>
<tr>
<td>New</td>
<td>21</td>
<td>33</td>
</tr>
<tr>
<td>Study</td>
<td>55</td>
<td>58</td>
</tr>
<tr>
<td>Series</td>
<td>80</td>
<td>74</td>
</tr>
<tr>
<td>Words</td>
<td></td>
<td></td>
</tr>
<tr>
<td>New</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>Study</td>
<td>87</td>
<td>88</td>
</tr>
<tr>
<td>Series</td>
<td>91</td>
<td>95</td>
</tr>
</tbody>
</table>

Note: Estimated standard errors of cell means are as follows: for digit-letter data, 1.6 within Feedback and 1.1 within Control; for word data, 5.6.

Results

Overall Effects of Base Rate, Feedback, and Stimulus Type

The basic data for analyses of the overall effects of the principal experimental factors are summarized in Tables 1 and 2 for Experiments 1 and 2, respectively, in terms of percentage of "old" responses averaged over stimulus orders and over the blocks of the test series. These percentages represent responses to items presented during the recognition test series, with the history of the items identified by the row heading in the Source column. Data in the New row are for "old" responses to items that occurred for the first and only time during the test series, in the Study row for items that had appeared in the study list but not previously during the test series, and in the Series row for items that had appeared (as new items) during the previous block of the test series.

The 33% and 67% column headings in Tables 1 and 2 identify the two base rates in terms of the true percentage of old (i.e., repeated) items in the test series. 4

Accuracy and bias. As has been the practice in nearly all research on recognition from the late 1960s to the present, we wish to distinguish effects of experimental factors on accuracy of recognition from effects on participants' criteria, or biases, for making "old" judgments. For this purpose, we have used the paired values for the New and Series rows in each column of Tables 1 and 2 to compute the measures of recognition accuracy and criterion presented in Tables 3 and 4, respectively. Henceforth, using the now standard terminology drawn from signal detection theory, we refer to "old" responses to new items as false alarms and to "old" responses to series items as hits. 4 In these terms, the familiar measure of accuracy, d', can be interpreted essentially as the difference between hits and false alarms converted to standard deviation units, with larger values of course signifying greater accuracy. Full discussions of this and other measures based on signal detection theory are given in Macmillan (1993), McNicol (1972), and Swets (1964). The only notable effect of our independent variables on accuracy was a strong and significant increase in d' from digit to letter stimuli and an even greater increase from letters to words. Feedback produced a small, marginally significant increase in accuracy over the control condition with letter and digit stimuli but no effect with words.

Table 2
Percentages of "Old" Responses to Each Stimulus Type, Averaged Over Order and Blocks (Experiment 2)

<table>
<thead>
<tr>
<th>Source</th>
<th>33% Base rate</th>
<th>67% Base rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Feedback</td>
<td>Control</td>
</tr>
<tr>
<td>Digits</td>
<td></td>
<td></td>
</tr>
<tr>
<td>New</td>
<td>31</td>
<td>57</td>
</tr>
<tr>
<td>Study</td>
<td>40</td>
<td>59</td>
</tr>
<tr>
<td>Series</td>
<td>68</td>
<td>72</td>
</tr>
<tr>
<td>Letters</td>
<td></td>
<td></td>
</tr>
<tr>
<td>New</td>
<td>28</td>
<td>36</td>
</tr>
<tr>
<td>Study</td>
<td>66</td>
<td>65</td>
</tr>
<tr>
<td>Series</td>
<td>73</td>
<td>70</td>
</tr>
<tr>
<td>Words</td>
<td></td>
<td></td>
</tr>
<tr>
<td>New</td>
<td>8</td>
<td>22</td>
</tr>
<tr>
<td>Study</td>
<td>80</td>
<td>68</td>
</tr>
<tr>
<td>Series</td>
<td>92</td>
<td>96</td>
</tr>
</tbody>
</table>

Note: Estimated standard errors of cell means are as follows: for digit-letter data, both feedback and control conditions, 4.8; for word data, 3.6.

Table 3
Average Accuracy (d') Values by Condition

<table>
<thead>
<tr>
<th>Experiment and source</th>
<th>Feedback</th>
<th>Control</th>
<th>Feedback</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digits</td>
<td>1.14</td>
<td>0.84</td>
<td>1.18</td>
<td>0.94</td>
</tr>
<tr>
<td>Letters</td>
<td>1.64</td>
<td>1.08</td>
<td>1.37</td>
<td>1.24</td>
</tr>
<tr>
<td>Words</td>
<td>2.57</td>
<td>2.77</td>
<td>2.88</td>
<td>2.70</td>
</tr>
<tr>
<td>Experiment 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digits</td>
<td>0.98</td>
<td>0.40</td>
<td>0.72</td>
<td>0.64</td>
</tr>
<tr>
<td>Letters</td>
<td>1.19</td>
<td>0.88</td>
<td>1.30</td>
<td>1.26</td>
</tr>
<tr>
<td>Words</td>
<td>2.80</td>
<td>2.56</td>
<td>2.92</td>
<td>2.68</td>
</tr>
</tbody>
</table>

4 Though participants could make judgments of either "old" or "new" concerning test stimuli, percentages of "old" and "new" judgments under any given condition are necessarily complementary, so throughout this article, we analyze and report only "old" percentages.

5 For the Experiment 1 digit-letter condition, data from the NFNB group and the NFB group agreed so closely that they have been combined and are presented henceforth under the label control (the Control heading in Tables 1 and 2).

6 Using the overall percentage of "old" responses to old stimuli as one of the quantities entering into computation of the d' measure would not do in this study because over test blocks it would represent a shifting mixture of items from the study and series categories.
independent of base rate with all stimulus types. The strong relation between accuracy and stimulus type is not surprising, but the explanation is not obvious because the stimulus types differed in many respects. Discussion of the roles of similarity and other stimulus attributes is deferred to the section on model-based analyses.

The measure of response bias we have chosen is an estimate of the observer’s criterion, C, for making ‘old’ responses, defined by

\[ C = -0.5(z_a + z_b) \]

where \( z_a \) and \( z_b \) denote hit and false-alarm rates, respectively, transformed to \( z \) scores (Macmillan, 1993). It has advantages over other proposed measures with respect to ease of computation and directness of its relation to hit and false-alarm rates. In Table 4, estimates of \( C \) (actually \( -C \), in order to produce mainly positive entries) are presented, computed from the hit and false-alarm values in Tables 1 and 2. A \( C \) value of 0 signifies no bias (i.e., equal biases toward ‘old’ and ‘new’ responses), and \( C \) values less than 0 or greater than 0, respectively, signify biases toward ‘old’ and ‘new’.

In all models deriving from signal detection theory, measures of an individual’s bias, or criterion, are assumed to reflect response tendencies that are independent of stimulus characteristics, so it is no surprise that, in contrast with the picture obtained for \( d' \), the bias estimates for our data exhibit no significant relation to stimulus type, the values in Table 4 being small in magnitude and varying inconsistently across stimulus types. The only appreciable and significant effect is a Feedback \( \times \) Base Rate interaction, with the addition of feedback to the standard procedure producing an increase in bias toward ‘old’ responses under the 67% base rate but a decrease under the 33% base rate.

To elucidate the nature of this interaction, we computed estimates of \( C \) for each 24-trial block during the test series. The results are presented in Figures 1 and 2 for the combined digit and letter conditions of Experiments 1 and 2, respectively. The curves for ‘learning’ under the control procedure, shown in the lower panels of Figures 1 and 2, exhibit differences between the 33% and 67% base-rate curves in the earliest blocks, which are presumably attributable to chance because the participants were not informed of the base rates in advance, but the differences disappear over later blocks as curves for the two base rates converge to a common level. A sharply different picture is presented by bias estimates for the feedback procedure, shown in the upper panels of Figures 1 and 2, with the curves for the 67% and 33% base rates diverging over blocks of the test series.

A statistical analysis of these results, based on \( d' \) estimates computed for individual participants, is summarized in Appendix A. Throughout the study, statistical significance was assessed by analyses of variance (ANOVAs) conducted in the framework of a mixed model (Arnold, 1981; Estes, 1991). For brevity, we refer to effects as significant or nonsignificant without qualification when tests yield \( p \leq .001 \) or \( p \geq .05 \), respectively.

---

**Table 4**

<table>
<thead>
<tr>
<th>Experiment and source</th>
<th>33% Feedback</th>
<th>33% Control</th>
<th>67% Feedback</th>
<th>67% Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXPERIMENT 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digits</td>
<td>.04</td>
<td>.22</td>
<td>.64</td>
<td>.24</td>
</tr>
<tr>
<td>Letters</td>
<td>.02</td>
<td>.10</td>
<td>.79</td>
<td>.15</td>
</tr>
<tr>
<td>Words</td>
<td>.06</td>
<td>.26</td>
<td>.31</td>
<td>.40</td>
</tr>
<tr>
<td>EXPERIMENT 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digits</td>
<td>-.01</td>
<td>.28</td>
<td>.86</td>
<td>.18</td>
</tr>
<tr>
<td>Letters</td>
<td>.01</td>
<td>.08</td>
<td>.39</td>
<td>.11</td>
</tr>
<tr>
<td>Words</td>
<td>.00</td>
<td>.49</td>
<td>.28</td>
<td>.54</td>
</tr>
</tbody>
</table>

*Note.* Entries are values of \( -C \), where \( C \) is the bias measure defined in the text.
trends that call for prediction or interpretation by recognition models.

We start with the results given in Tables 1 and 2 for the control procedure, which has been the standard procedure in the vast majority of preceding studies. The "old" percentages in the rows labeled New (or false-alarm rates in conventional terminology) depended strongly on stimulus type, increasing uniformly from words to letters to digits in both experiments, but were independent of old stimulus base rates, exhibiting only very small and unsystematic differences between the 33% and 67% conditions. Correct response rates, that is, "old" percentages in the Study and Series rows, depended less strongly on stimulus type but uniformly lined up in an inverse order, increasing from digits to letters to words, and exhibited a small but fairly uniform tendency to increase from the 33% to the 67% base rate.

Stimulus effects were essentially the same under the feedback procedure, but, in contrast to the control results, false-

Figure 2. Estimates of the bias measure \(-C\) by block for 67% and 33% base-rate conditions of Experiment 2. Estimates are based on data averaged over digit and letter stimulus types.

Bias estimates for the word conditions are shown in Figures 3 and 4 for Experiments 1 and 2, respectively. Again, under the control procedure, curves for the 33% and 67% base rates tend to converge to a common level over the test series. Feedback exerts little or no effect on bias in the word conditions, however, with none being apparent in the top panel of Figure 3 and only a faint suggestion of a late divergence of the 33% and 67% base-rate curves being discernible in the top panel of Figure 4. It is apparent that for all stimulus types, in the absence of feedback, people do not learn to adjust their response biases to reflect old–new base rates.

"Old" response rates by stimulus source. The basic data from our experiments are not derived measures of accuracy or bias, but percentages of "old" judgments, and we are concerned with the latter in the remainder of our presentation of results and in all model-based analyses. Here we summarize, with statistical support as appropriate, some major effects and

Figure 3. Estimates of the bias measure \(-C\) by block for word stimuli in the 67% and 33% base-rate conditions of Experiment 1.
alarm rates and both types of correct response rates increased uniformly from the 33% to the 67% base-rate condition, the effects being small for word stimuli but large for letter and digit stimuli. All of these trends may be more easily seen in Figures 5 and 6 for the control condition and Figures 7 and 8 for the feedback condition.

This pattern of effects was substantiated by statistical analyses. In the letter and digit conditions of both experiments, the effect of stimulus type and the Base Rate × Feedback-Control interaction were significant. Within the feedback procedure, the base-rate effect was significant, but within the control procedure it was nonsignificant. In the word conditions of both experiments, there were no significant base-rate or feedback effects. For all stimulus types, differences among the percentages of "old" responses on new, study, and series items were significant (both by overall F tests and by pairwise contrasts).

![Figure 4](image)

*Figure 4.* Estimates of the bias measure $-\alpha$ by block for word stimuli in the 67% and 33% base-rate conditions of Experiment 2.

![Figure 5](image)

*Figure 5.* Percentages of "old" responses to new items (New), items repeated from the study series (St.), and items repeated from the preceding test block (Ser.) by stimulus type for the combined control condition of Experiment 1. The 33% and 67% base-rate conditions are labeled BR 33 and BR 67, respectively. Predicted (Th.) and observed (Obs.) percentages are averaged over blocks.

**Trends Over Test Blocks**

Trends over blocks of the recognition test series were very similar for letter and digit stimuli, so, for stability, averages for these conditions are shown in Figures 9 and 10 for Experi-

---

8 Test statistics and $p$ values are summarized in Appendix B.
Figure 6. Percentages of "old" responses to new items (New), items repeated from the study series (St.), and items repeated from the preceding test block (Ser.) by stimulus type for the control condition of Experiment 2. The 33% and 67% base-rate conditions are labeled BR 33 and BR 67, respectively. Predicted (Th.) and observed (Obs.) percentages are averaged over blocks.

Figure 7. Percentages of "old" responses to new items (New), items repeated from the study series (St.), and items repeated from the preceding test block (Ser.), computed by stimulus type for the feedback condition of Experiment 1. The 33% and 67% base-rate conditions are labeled BR 33 and BR 67, respectively. Predicted (Th.) and observed (Obs.) percentages are averaged over blocks.
extended test series of Experiment 2. For word stimuli (upper panels of Figures 15 and 16), however, FA rates for the 33% and 67% base rates exhibit only a very slight, and statistically insignificant, divergence over the test series, and there is no hint that the FA rates would ever approach the base rates.

**Model-Based Analyses**

**Summary of the basic array model.** The array model derives from the family of categorization models originated by Medin and Schaffer (1978) and known in the literature as context models (Nosofsky, 1984, 1986, 1988) or exemplar models (Estes, 1986a; Medin & Florian, 1992). We use the latter designation. Exemplar models have in common the assumption that representations of the stimuli (category exemplars) encountered by a learner in a categorization task and encoded in terms of their attributes or features are stored in a memory array. It is convenient to speak of the stimuli that have been indicated to belong to a Category $A_i$ when they were presented during learning as constituting the Category $A_i$ subarray. When a learner is presented with a test stimulus, whether new or old, he or she compares its similarity to the stored exemplars of each category and makes a categorization judgment based on the relative total similarities of the test exemplar to the category subarrays. Several investigators have noted that the model can be readily extended to the interpretation of recognition on the assumption that a recognition judgment concerning a test stimulus is based on its total similarity to the memory array for all stimuli encountered during a preceding study or recognition-test series (Estes, 1986b; Medin & Schaffer, 1978; Nosofsky, 1988).

We illustrate the application of the array model to recognition in terms of the following example, in which a sequence of items, $I_1, I_2, \ldots, I_n$, is presented to a learner. We assume that, in response to the task instructions, the learner sets up old and new memory arrays and that representations of all items presented are stored in the old array. After a few trials, in which an item $I_i$ occurs twice and several others occur once each, the arrays take the form

<table>
<thead>
<tr>
<th>Old</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_1$</td>
<td>$I_0$</td>
</tr>
<tr>
<td>$I_2$</td>
<td></td>
</tr>
<tr>
<td>$I_3$</td>
<td></td>
</tr>
<tr>
<td>$I_4$</td>
<td></td>
</tr>
</tbody>
</table>
The sole entry I₀ in the new array does not correspond to any presented item. It represents, rather, the learner’s residual memory of any preexperimental experiences with stimuli of the same types as those that occur in the experiment. It is assumed, intuitively speaking, that a learner’s likelihood of recognizing an old test item depends on how much the information about the item retained in the old array exceeds the background level denoted by I₀. If, following this sequence, item I₁ occurs on a recognition test, its similarities to the old and new memory arrays are computed by summing its similarities to the individual representations in each array. The similarity of the item to the old array, so computed, is denoted Sim(Old) and its similarity to the new array, Sim(New). We assume that items in the population sampled have average similarity s₀ to I₀, so Sim(New) is simply equal to s₀. The probability of an “old” response on this test of I₁ is obtained by entering the computed similarities in the general formula derived from exemplar models,

$$P(\text{Old}) = \frac{\text{Sim}(\text{Old})}{\text{Sim}(\text{Old}) + \text{Sim}(\text{New})}.$$  \hspace{1cm} (2)

Before proceeding with further description and illustration of the array model, we need to say a word about stimulus representation. The general assumption is that any stimulus is encoded in memory in terms of its features, or values on a list of relevant attributes. The attribute structure may or may not be known to the investigator. When it is not, or when for some

---

Figure 10. Predicted (Th.) and observed (Obs.) trends over blocks for “old” responses to new, study, and series items in the control condition of Experiment 2. Values plotted are averages over digit and letter stimulus types. s denotes the similarity parameter; Av. denotes average.

Figure 11. Predicted (Th.) and observed (Obs.) trends over blocks for “old” responses to new, study, and series items for word stimuli in the control condition of Experiment 1.

Figure 12. Predicted (Th.) and observed (Obs.) trends over blocks for “old” responses to new, study, and series items for word stimuli in the control condition of Experiment 2.
reason one wishes to generate predictions that do not depend on details of stimulus structure, one may use a form of the model, henceforth denoted the $s_{\text{pattern}}$ version, in which only global measures of similarity between stimulus patterns enter into the computations.

We illustrate the $s_{\text{pattern}}$ version of the model in terms of the current example, starting with the similarity computation for Test Item $I_t$. The $I_t$ pattern is compared with each of the representations in the old array. When the test item matches a representation in memory, a value of 1 is entered in the sum that will constitute $\text{Sim}(\text{Old})$. When a mismatch occurs, a quantity $s$ is entered, the "similarity parameter" $s$ having a value between 0 and 1 and representing the average interitem similarity for the population of items being sampled. Thus, the similarity of test item $I_t$ to the old array, $\text{Sim}(\text{Old})$, is $1 + s + s + 1 + s = 2 + 3s$, which in this example yields

$$P(\text{Old}) = (2 + 3s)/(2 + 3s + s_0).$$  

(3)

If, instead, a new item, say $I_{t+1}$, were tested, the probability of incorrectly calling it old, that is, producing a false alarm, would be

$$P(\text{Old}) = 5s/(5s + s_0).$$  

(4)

These expressions can easily be generalized. If we denote the total number of items in the sequence presented to the learner by $N$ and the number of repetitions of item $I_t$ by $k$, the probability of recognizing item $I_t$ becomes

$$P(\text{Old}) = [k + (N-k)s]/[k + (N-k)s + s_0]$$  

(5)
alternative form, which we term the $s_{\text{pattern}}$ version, we need to introduce some assumptions about stimulus structure. Suppose that in the example the items are objects describable as either black (b) or white (w) and either circular (c) or triangular (t), and that the item representations are $I_1$: bc; $I_2$: bt; $I_3$: wc; and $I_4$: wt, that is, $I_1$ is a black circle, and so on. The memory array now takes the form

\[
\begin{array}{cccc}
\text{Old} & \text{New} \\
bc & I_0 \\
bt & \\
w & c \\
bc & w \\
wt & .
\end{array}
\]

In this version of the model, the similarity parameter, $s$, is assumed to represent average similarity between nonidentical

Figure 15. Predicted (Th.) and observed (Obs.) trends over blocks for "old" responses to new, study, and series items for word stimuli in the feedback condition of Experiment 1. The 33\% and 67\% base-rate conditions are shown in panels labeled BR 33 and BR 67, respectively.

and the probability of incorrectly calling a new item old becomes

\[
P(\text{Old}) = N_s / (N_s + \alpha_0). \tag{6}
\]

Except in the degenerate case when $s = 1$ (all items indistinguishable), the probability of a correct recognition response to an old item will be higher than the probability of a false alarm to a new item. Several other predictions are immediately obvious: For example, provided only that $s$ is greater than 0, the probability of recognizing an item increases as a function of the number of times it is repeated, and probabilities of both hits and false alarms increase as a function of the total number of items that have been presented.

To show how the $s_{\text{pattern}}$ version of the model differs from the

Figure 16. Predicted (Th.) and observed (Obs.) trends over blocks for "old" responses to new, study, and series items for word stimuli in the feedback condition of Experiment 2. The 33\% and 67\% base-rate conditions are shown in panels labeled BR 33 and BR 67, respectively.
features. Thus, on a test of bc, the similarity of the test item to its own representations is 1, to bt is s, to wc is j, and to w is $s^2$. Yielding \( \text{Sim} (\text{Old}) = 1 + s + s + 1 + s^2 = 2 + 2s + s^2 \).

Obviously, the \( s_{\text{feature}} \) version of the model will be much more sensitive than the \( s_{\text{pattern}} \) version to the particular sequence of items presented to the learner and to the particular test items used.

In all fits of the array model to our data that are reported in this article, we used the \( s_{\text{feature}} \) version except when the two versions were being explicitly compared. For digit and letter stimuli, the featural similarity parameter was assumed to index the similarity between any two nonidentical digits or nonidentical letters in the position of the digit or letter triad being compared. For word stimuli, matters were more complex, for there is no general agreement on what constitutes features of words. What we did, for comparability, was to apply the parameter \( s \) to the first two letters of words that were being compared exactly as done for letter triads and to define an artificial third feature comprising the remainder of the letters of each word taken together, with an associated similarity parameter denoted \( s_9 \). The parameter \( s \) enters into similarity computations just as does \( s \). For example, because the words \textsc{shaker} and \textsc{snake} match on their first letter, differ on their second letter, and differ with respect to their artificial third feature (\textsc{aker} vs. \textsc{ake}), their similarity is computed to be \( 1 \cdot s \cdot s \). Estimates of \( s \) prove to be smaller in magnitude than those of \( s \), so even two words with identical first and second letters (e.g., \textsc{shif} and \textsc{shim}), whose similarity is \( 1 \cdot 1 \cdot s \), may have very small similarities.

Only one important constituent of the model remains to be described, namely provision for retention loss. For reasons that have been described in detail elsewhere (Estes, 1993, 1994; see also Nosofsky et al., 1992), we assumed that the level of availability of the memory representation of an item is reduced by a factor \( \alpha \) during each subsequent item presentation. The "decay parameter" \( \alpha \) has a value between 0 and 1. With this parameter included in the model, when the similarity of a test item to the old memory array is computed, each constituent similarity that enters into the sum is weighted by the appropriate power of \( \alpha \). Thus, for the \( s_{\text{pattern}} \) version of the model, the similarity of test item \( t_1 \) to the old array in the example becomes \( \text{Sim}(\text{Old}) = \alpha t + \alpha s + s + \alpha + s \), which of course reduces to Equation 3 if \( \alpha \) is equal to 1. Analogously, similarity of a new test item to the old array becomes \( \alpha t + \alpha s + \alpha t + \alpha s + s = s(1 - \alpha^2)/(1 - \alpha) \), and Equation 6 is modified to the more general form

\[
P(\text{Old}) = N's/(N's + s_0),
\]

where \( N' = (1 - \alpha^2)/(1 - \alpha) \).

It is desirable under some circumstances to also be able to represent response bias in the model. To do so, we introduced a parameter, \( b \), with a value between 0 and 1, representing bias toward choosing the "old" response independently of stimulus properties (following Nosofsky, 1988). This parameter enters into the model by way of a simple modification of the model's basic expression for response probability, Equation 2, which becomes

\[
P(\text{Old}) = b\text{Sim}(\text{Old})/[b\text{Sim}(\text{Old}) + (1 - b)\text{Sim}(\text{New})].
\]

Replacing \( \text{Sim}(\text{New}) \) by the background parameter, \( s_9 \) as done in Equations 3–7, we have

\[
P(\text{Old}) = b\text{Sim}(\text{Old})/[b\text{Sim}(\text{Old}) + (1 - b)s_9],
\]

which can be rearranged in the form

\[
P(\text{Old}) = \text{Sim}(\text{Old})/[\text{Sim}(\text{Old}) + s_9^*],
\]

where

\[
s_9^* = (1 - b)s_9/b.
\]

The expression for \( P(\text{Old}) \) takes the same form for tests of either old or new items, but of course the value of \( \text{Sim}(\text{Old}) \) will systematically differ, always being larger at a given point in the test series for a test on an old item. It is easy to see that we have not changed the number of free parameters of the model, for wherever \( s_9 \) appears in the original model, it will be replaced by \( s_9^* \) in the generalized model, so the number of parameters to be estimated from data is not increased. However, the value of the composite parameter \( s_9^* \) is assumed to depend on the learner's response bias, if any, as well as on similarity of test items to the background information incorporated in the term \( s_9 \) of the basic array model.10

In our applications of the array model to the data of this study, we used the denotation \( s_9 \) for the background parameter in analyses of the control condition but had occasion to use the more general denotation \( s_9^* \) in analyses of the feedback condition.

Our standard procedure for applying the model to a data set is to use a hill-climbing computer program to estimate the model's parameters. The program finds the values that minimize the sum over the test series of squared differences between theoretical and observed response proportions for individual test trials. The parameters are estimated separately for different stimulus types and different feedback conditions. However, it is important to note that, in all instances, a single set of estimates of the parameters is computed for the combined data from the 33% and 67% base-rate procedures within any Stimulus Type \( \times \) Feedback Condition combination. Thus, relations between theoretical values for response mea-

9 We do not regard this strategem as very satisfactory, and the \( s_{\text{feature}} \) version of the array model did not fit our word data much better than did the \( s_{\text{pattern}} \) version. It seems likely that improving the fit significantly would require taking account of semantic attributes of words.

10 We do not have occasion to use the receiver operating characteristic (ROC) functions in this study. If one wished to do so in other applications of the model, one would plot pairs of theoretical values of hits and false alarms with similarity of a test stimulus to the old memory array, \( \text{Sim}(\text{Old}) \) in Equation 2, held constant while the value of the parameter \( s_9^* \) was allowed to vary, just as an ROC curve is obtained in studies of signal detectability by holding the discriminability parameter \( d' \) of signal detection theory constant while the observer's criterion is allowed to vary.
sures obtained under 33% versus 67% base rates in the control condition are genuine predictions from the model.

In presenting our model-based analyses, we take up the control and feedback conditions of the two experiments successively because the three-parameter array model proved satisfactory for interpretation of the control conditions but included no way of representing the difference between feedback and no feedback. Thus, when we take up the feedback conditions, we augment the model with a process that entails estimating one additional parameter.

**Control condition.** The array model predicted the absence of base-rate effects observed in our control conditions independently of parameter values. The basis of the prediction is easiest to see for false alarms, because a new test item has the same average similarity to all item representations in the old memory array, regardless of the relative frequencies with which those items were presented. Thus, the general, a priori prediction concerning the trend in false-alarm probabilities over test trials is as shown in Figure 17, the upper panel showing curves computed from Equation 6 (with the decay parameter, **α**, set equal to 1, for no decay) and the lower panel showing curves computed from Equation 7 (with **α** set equal to .9999). The steepness with which the false-alarm rate rises over trials is strongly determined by interitem similarity, and even the small rate of decay represented by the reduction of **α** from 1.0 to .9999 between the upper and lower panels of Figure 17 modifies final levels as well as the steepness of the false-alarm functions.

Thus, the array model implies that probabilities of old responses, not only to new items but also to items repeated from the study list or the earlier trials of the test series, should be independent of base rates under the control procedure. This implication arises because for either type of old test item, similarity to the old memory array is simply its similarity to its own representation in the array plus its summed similarity to all other items in the array, which depends only on stimulus properties of the items and not on their presentation histories.

Predictions computed from the model for average percentages of old response proportions over the test series for each stimulus type are included together with the corresponding data values in Figures 5 and 6 for Experiments 1 and 2, respectively. Both the way the level of old responding increased from new to study to series and the invariance over base rate within each of these categories seemed to be well predicted by the model.

The other conspicuous feature of Figures 5 and 6—the strong and uniform increase in level of false alarms from words to letters, digits, coupled with parallel decreases in levels for both of the old categories—was well described by the array model but could not be said to be predicted a priori because the parameters were estimated separately for each stimulus type. It is of interest, however, to look at the parameter estimates for the light they may shed on the factors contributing to the trends across stimulus type. For this purpose, we computed averages over Experiments 1 and 2 of the parameter estimates that yielded the model fits shown in Figures 5 and 6. Considering first the background constant **β** (i.e., the parameter most closely reflecting criterion, or bias), the values obtained were .48, .35, and .08 for the digit, letter, and word conditions, respectively, implying a progressively increasing tendency to call stimuli old.11

![Array Model with No Decay](image)

![Array Model with Decay](image)

**Figure 17.** A priori predictions of the array model with no memory decay (upper panel) and with decay (lower panel) for trends of false alarm (FA) rates over blocks. The predicted functions are independent of old-new stimulus base rate but sensitive to interitem similarity (**s**).

11 The reason is that smaller values of **β** in Equation 5 or 6 yield larger values of **P(Old)**.
version of the model would not be expected to behave analogously because it reflects only degrees of similarity between different individual digits or different individual letters. In our fits of the \( s_{\text{stimulus}} \) version, the estimate of \( s_{\text{stimulus}} \) was .04 for both letter and word stimuli (which is reasonable enough because it reflected similarities between letters in both cases) and .06 for digits.

Differences among stimulus types with respect to the decay parameter are also to be expected. One reason for this expectation is that more wordlike stimuli are more rehearsable. A second reason is that for the different stimulus types, items intervening between study and test vary in average similarity to the test item. The mean estimates of the decay parameter (.886, .994, and .995 for digits, letters, and words, respectively) do not look large, but they imply much greater decay over the test series for digit stimuli than for the other types. Over the course of 100 trials, for example, availability of a word stimulus would decline from its initial level of 1.0 to .602, a letter stimulus to .548, and a digit stimulus to .00001.

To address the more difficult task of predicting trends over the recognition test series, we computed the theoretical values for block-by-block response proportions shown in Figures 9, 10, 11, and 12 using the same parameter estimates as in Figures 5 and 6. Figures 9 and 10 present data and predictions averaged over digit and letter stimulus types as well as base rates; Figures 11 and 12 present corresponding data and predictions for the word conditions. The model fits shown in Figures 9, 10 (upper panel), 11, and 12 are for the \( s_{\text{stimulus}} \) version of the array model. In addition, a comparison of the two versions of the model is provided in Figure 10, where the lower panel shows predictions from the \( s_{\text{stimulus}} \) version. The predictions from the \( s_{\text{stimulus}} \) version follow the irregularities of the observed functions more closely, the most dramatic instance being the large decrease in old responding to study items coupled with increases to new and series items in Block 8.

Several aspects of our theoretical account of the data need comment. The predictions about the absence of base-rate effects and the qualitative pattern of trends over test blocks for the new, study, and series item categories are firm implications of the model that do not depend on estimates of parameters from the data. Predictions about effects of stimulus type have a somewhat different status, for the similarity parameter, \( s \), the background constant, \( s_{\text{bg}} \), and the decay parameter, \( \alpha \), were estimated separately for each stimulus type. When designing the experiments, we assumed on the basis of previous research (Estes, 1994) that interitem similarities would be largest for digits, considerably smaller for consonant strings, and still smaller for words, and the estimates of \( s \) computed once the data were in hand agreed with these expectations. These interitem similarity differences contributed strongly to the predictions of the model regarding effects of stimulus type. The effect of similarity appears again in the role of the background parameter \( s_{\text{bg}} \), estimates of which line up with stimulus type in the same order as estimates of \( s \). Evidently, on average, test stimuli of the digit category are most similar and test stimuli of the word category least similar to memory representations of stimuli of the same type that were encountered in a participant’s preexperimental experience.

**Feedback condition.** The pattern of effects of stimulus type and base rate on overall levels of “old” responding to new, study, and series items under the feedback procedure is similar to the pattern observed in the control data in some respects but divergent in others. The percentages of “old” responses by source of item for the feedback data, shown in Figures 7 and 8, like those for the control data (Figures 5 and 6), exhibit a strongly increasing trend in false alarms from word to letter to digit stimuli within each base-rate condition, coupled with decreasing trends for “old” responses to study and series items. However, comparison of the results for 33% and 67% base rates in each of the bar diagrams of Figures 7 and 8 shows large increases in false alarms from the 33% to the 67% condition for both digit and letter stimuli, in sharp contrast to the absence of base-rate effects seen for the control data in Figures 5 and 6. Thus, a new problem for the array model, or any other recognition model, is to account for the observed effect of feedback on false-alarm rates with both digit and letter stimuli. A companion problem is to account for the absence of this effect with word stimuli.

The array model, as developed up to this point, like other extant recognition models, lacks any conceptual machinery for representing feedback effects, but we have considered two possible augmentations of the model that might remedy matters. In view of our findings about trends in response bias under feedback (Figures 1-4), the most obvious route is to assume that bias toward “old” or “new” responses is driven by feedback. Our previous analyses indicate that, under the control procedure, response bias does not vary significantly over the test series and, in particular, does not depend on the prevailing base rate over the range we investigated. These invariances should, of course, hold for the value of the bias parameter \( b \) defined in Equation 8. Under the feedback procedure, however, where response bias does depend on base rate, we assume that the value of \( b \) increases on any test trial when feedback indicates that the tested item was old and decreases when feedback indicates that the item was new. The result is that under the 67% base rate, when old trials are more frequent, \( b \) will increase over test trials, causing the false-alarm rate to increase, and under the 33% base rate, when new trials are more frequent, \( b \) will decrease, causing the false-alarm rate to decrease. Details of the assumed learning process are described in Appendix C.

The alternative augmentation of the array model that we have investigated stems from the observation that if participants did not attend to the items presented, the recognition test series under the feedback procedure would have been closely analogous to a traditional probability learning experiment (Humphreys, 1939). That is, participants would simply be learning to predict the occurrence of old versus new feedback signals in the background context present on all trials. Thus, we propose that under the feedback procedure, a contextual learning process proceeds in parallel with the array model process, and recognition performance is generated by the outputs of these processes in a stepwise fashion. The learner first consults the output of the array model process, and if it leads to an “old” response, the trial ends; if not, the learner responds in accord with the output of the contextual learning
process. Details of the assumed learning process are given in Appendix C.

An interesting property of both versions of the augmented model is the common prediction that false-alarm probability should not approach the old stimulus base rate asymptotically, but rather (in the absence of memory decay) should approach unity. However, quantitative predictions from the two versions concerning the form of the false-alarm function differ enough that it is possible in principle to choose between them on the basis of their accounts of the data from the feedback condition. The present experiments were not designed for this purpose, however, and the results are not decisive. Fits of the two augmented models to our data yielded a small but consistent advantage for the contextual learning over the bias adjustment version, as shown in Appendix C, so in the remainder of this analysis of the feedback condition, we present only fits of the contextual learning version.

Because the augmented model includes the basic array model as a component, predictions about the general pattern of results in the feedback condition must be similar to those for the control condition except for the appearance of base-rate effects. Differences in base rate are predicted by both augmented models to affect performance on all three types of items (new, study, and series), with the most conspicuous effect being a tendency for false-alarm rates to move toward old item base rates over test blocks. These expectations are borne out by the analyses presented in Figures 13–16. These figures parallel Figures 9–12 for the control condition, except that, with the addition of feedback, it is not appropriate to pool data over base-rate conditions. Because the results for the 33% and 67% base rates have to be presented separately, the empirical functions are necessarily less stable than those for the control condition, and we cannot expect fits of the model to the data to be as good. Nonetheless, the pattern of results for overall mean percentages of “old” responses to new, study, and series items under both base rates (Figures 7–8 for Experiments 1 and 2, respectively) are quite well predicted by the model. In particular, the substantial observed increases in false-alarm rates from the 33% to the 67% base-rate condition with both digit and letter stimuli and the virtual absence of this effect with word stimuli are closely reflected in the theoretical values computed from the model. The observed functions for performance over test blocks (Figures 13–16) are quite irregular, and only the qualitative trends can be said to be predicted by the model.

The virtual absence of a base-rate effect with word stimuli is accommodated by the model only because the estimated rate of the contextual learning process (indexed by the parameter b in Appendix C, Equation C3) proves to be very small. The problem remains, however, of explaining why the process of acquiring information about base rate should be so much slower with words than with other types of stimuli.

General Discussion

Summary of Findings and Interpretations

Our principal results bear on the effects and interactions of stimulus attributes, old–new base rates, and feedback conditions. This study goes beyond other contemporary work on recognition both in extending the range of stimulus types investigated and in providing controlled comparisons of the effects of stimulus variables at different levels of other factors. The most ubiquitous and robust finding was the strong dependence of all response measures on stimulus type. Rates of false alarms to new stimuli increased from words to letters to digits, whereas rates of correct responding to old stimuli—both those included in the study list and tested for the first time at varying lags in the recognition test series and those first presented in the test series and tested at constant average lags—decreased from words to letters to digits, and these trends held for every combination of the base rate and feedback conditions. In terms of the array model, perhaps the most important factor in these trends was the systematic increase in interitem similarity from words to letters to digits. This factor was manifested directly in the trend for false-alarm rates: Higher interitem similarities produced higher global similarities of new test items to the stored representations of old items and, therefore, more false alarms. However, similarity must also be assumed to be implicated in the increasing estimated decay rates from words to letters to digits, for it is well established that retention loss between study and test increases with similarity of intervening items (Postman, 1971).

It should be noted that the term similarity, as used here, is defined by the functional relations it enters into, as those expressed in Equations 3–7, and the measure of similarity assumed is the value of the similarity parameter, s, estimated from recognition data. We would intuitively expect similarity, so defined, to be related to similarity as assessed by similarity judgments (as in the work of Tversky, 1977), but the nature of the relationship is an open research question. The most important point for present purposes is that similarity as defined in the array model framework is assumed to reflect confusability between stimulus representations, whereas similarity judgments may have no simple relation to confusability. We find words to have very low confusabilities relative to some other kinds of stimuli, notably digit and consonant strings, commonly used in recognition research and suggest that this factor may be related to the fact that some phenomena conspicuous with other types of stimuli (e.g., the interaction of base rate with feedback conditions) are absent with word stimuli.

A finding that would not have been anticipated on the basis of earlier literature (Donaldson & Murdock, 1968; Olson, 1969; Parks, 1966) was the absence of any significant effect of old–new base rate on false-alarm frequency when recognition testing was carried out without feedback about correctness of responses (our control condition). This result was predicted a priori by the array model on the premise that participants would receive information about base rates only by way of their experience during study and testing. This qualification is important, because Ratcliff et al. (1992) found a direct relation between base rate and false alarms in a study in which participants were given advance information about base rates prior to each recognition test series. In terms of the array model, the prior information may have enabled participants to adjust their response biases so as to reflect base rates, thus modifying the parameter b (defined in Equation 8) and thereby their predicted false-alarm rates as well.
Although it appears that false-alarm rate is independent of old–new base rate under the standard procedures of most recognition research, we did find a dependence, evident only for the digit and letter stimulus types, when we added informative feedback during recognition testing. In an attempt to develop a quantitative account of the observed trends in responding under feedback, we considered two versions of a combined model in which information storage and similarity computations proceed exactly as in the basic array model but coupled with a feedback-driven learning process. In one version, it is assumed that, in parallel with the array model process, the participant acquires information about probabilities of old and new feedback signals in the given experimental context. In the other version, the basic array model is modified only by the assumption that the magnitude of the bias parameter (the parameter $b$ in Equations 8 and 9) increases whenever feedback indicates "old" to be the correct response on a test trial and decreases whenever feedback indicates "new" to be correct. Both versions account reasonably well for the qualitative trends in responding in the feedback condition. The contextual learning version proved somewhat superior quantitatively (see Appendix C), though less than satisfactory. Both versions appear to merit further investigation.

**Comparison of Alternative Models**

In this section, we consider how our results on some relatively neglected variables can be addressed by currently influential models of recognition. We limit our attention to three that are quantitatively formulated and comparable in scope to the array model—MINERVA (Hintzman, 1988), the search of associative memory (SAM) model (Gillund & Shiffrin, 1984; Shiffrin et al., 1990), and the theory of distributed associative memory (TODAM) model (Murdock, 1982, 1991; Murdock & Lason, 1988).

All of these models are based on the idea that recognition depends on the relation between a perceived stimulus pattern and a composite memory. Assumptions about memory storage formats vary. In SAM, there are no specific assumptions about stimulus coding; an image of each item studied in a recognition experiment is stored with some probability, and on each repetition of an item, the strength of association of its image to the context and to other images, including itself, is increased. Perception of a test item produces a heightened level of activation of the entire composite memory by virtue of the associative connections of its image to the stored images. In MINERVA and TODAM, as in the array model, items are encoded as vectors (lists) of feature or attribute values, each repetition of an item produces a new stored vector, and a test item activates the composite memory by virtue of its similarities to the stored vectors.

All of these models embody a property of relative in that the level of activation of the composite memory by a test item does not determine recognition directly, but, rather, by way of a comparison with some reference quantity. In both MINERVA and SAM, total memory activation is compared with a criterion set by the participant, with a positive recognition (an "old" response in our experiments) occurring if the activation level exceeds the criterion. Thus, the formal role of the criterion is similar to that of the background-similarity constant $a_0$ of the array model. In TODAM, the total similarity of a test item to composite memory is compared with a measure of variability of the levels of activation of memory produced by new items in order to generate a theoretical index of recognition in terms of the $d'$ measure of signal detection theory.

Turning to the findings of this study, the most robust and ubiquitous was the increase in probabilities of both hits and false alarms over the recognition test series, with the curve for hits rising more rapidly at first but the two curves converging so that $d'$ tended to decrease on later trials. In terms of all four of the models, the increases are due to the growth of the composite memory over trials and the consequent higher levels of total activation generated by test items. The function for hits is always higher than that for false alarms because an old test item necessarily has at least one (in our study, exactly one) stored representation of itself in memory, whereas a new item has none. For the SAM model, total activation of memory is higher for an old item because the strength of its association with its own image is larger than the strength of association of a new item to any stored image. For the other three models, the similarity of an old item to its stored representation is greater than the similarity of a new item to any stored representation. Obviously, under a procedure that allows only one repetition of an old item, this advantage for old over new items will decrease in relative terms as the composite memory grows, yielding the predicted convergence of hit and false-alarm functions over trials. We have seen that the array model yields a quantitative account of these functional relations. It may well be that some or all of the other models could do the same, but it is not clear to us that they could do so with as few free parameters estimated from the data.

Our finding of invariance of false-alarm rates over old–new item base rates in the absence of feedback or prior information about base rate has been seen to be a direct, parameter-free prediction of the array model. The finding is not necessarily incompatible with the other models, but it does not seem to be predicted by them in a comparably direct way. We are not in a position to fit the MINERVA, SAM, or TODAM models to our data, and we are not sure, on the basis of the published descriptions of the models, whether predictions about base-rate effects would depend on assumptions about how participants set their criteria for "old" responses or on estimates of variances of "noise" distributions (e.g., distributions of activation levels generated by new test items).

The relation between recognition response tendencies and feedback in the form of information about correctness or payoff values is not represented in any of the models considered. It has proved possible to combine the basic array model with either a bias adjustment process or a contextual learning process so that, in either case, the resulting "mixed" model yields a fairly satisfactory account of our observed feedback effects at the cost of only one added parameter estimated from the data.

We note that our method of combining a composite memory model with a contextual learning process is quite general and could be used to augment the MINERVA, SAM, or TODAM models in a similar way for purposes of handling feedback effects.
References


Appendix A

Statistical Analyses of Accuracy (d') and Bias (C) Measures

Accuracy

Estimates of d' can be obtained either by entering the average hit and false-alarm proportions in d' tables (Swets, 1964), as done in preparing Table 3, or by entering the hit and false-alarm proportions of individual participants in the tables and averaging the individual estimates. The first method has some advantage with respect to stability (Macmillan & Kaplan, 1985), but the second method is necessary when statistical tests are desired. The averages of individual estimates for our data were slightly larger in absolute magnitude than the values shown in Table 3, as would be expected, but both sets of estimates agreed with respect to the ordering of conditions.

For Experiment 1, an ANOVA computed on the d's for the digit-letter condition yielded a significant effect only for stimulus type, \( F(1, 32) = 15.4, p < .001, MSE = .086 \). The feedback effect was marginal, \( F(1, 32) = 4.63, p < .05, MSE = .345 \). For Experiment 2, digit-letter condition, stimulus type yielded a significant effect, \( F(1, 40) = 18.9, p < .001, MSE = .168 \), and the feedback effect was short of significance, \( F(1, 40) = 2.95, p < .10, MSE = .168 \). For both experiments, all tests for the word conditions yielded Fs less than unity.

Bias

Analyses parallel to those for d' were computed on the bias measure C. For Experiment 1, stimulus type and feedback were nonsignificant, but base rate yielded a significant effect, \( F(1, 24) = 9.34, p < .01 \), as did the Feedback x Base Rate interaction, \( F(2, 24) = 3.35, p = .01, MSE = .17 \). For Experiment 2, the effect of stimulus type was marginal, \( F(1, 40) = 4.51, p < .05 \), and the effect of feedback was nonsignificant. Base rate yielded a significant effect, \( F(1, 40) = 7.49, p < .01 \), and a significant Feedback x Base Rate interaction, \( F(1, 40) = 14.28, p = .001, MSE = .10 \).

Appendix B

Analysis of Variance Test Statistics for "Old" Percentages

<table>
<thead>
<tr>
<th>Condition, variable, and experiment</th>
<th>( F )</th>
<th>df</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Letter-digit</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source (new, study, series) 1</td>
<td>283.5***</td>
<td>2, 48</td>
<td>.01</td>
</tr>
<tr>
<td>Source 2</td>
<td>186.1***</td>
<td>2, 64</td>
<td>.007</td>
</tr>
<tr>
<td>Source x Stimulus Type 1</td>
<td>21.1***</td>
<td>2, 48</td>
<td>.004</td>
</tr>
<tr>
<td>Source x Stimulus Type 2</td>
<td>29.9***</td>
<td>2, 64</td>
<td>.007</td>
</tr>
<tr>
<td>Feedback x Base Rate 1</td>
<td>7.6**</td>
<td>2, 24</td>
<td>.035</td>
</tr>
<tr>
<td>Feedback x Base Rate 2</td>
<td>12.2***</td>
<td>1, 32</td>
<td>.04</td>
</tr>
<tr>
<td><strong>FA averaged over blocks</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stimulus type 1</td>
<td>19.0***</td>
<td>1, 24</td>
<td>.004</td>
</tr>
<tr>
<td>Stimulus type 2</td>
<td>20.4***</td>
<td>1, 32</td>
<td>.261</td>
</tr>
<tr>
<td>Base rate 1</td>
<td>8.0**</td>
<td>1, 24</td>
<td>.02</td>
</tr>
<tr>
<td>Base rate 2</td>
<td>5.8*</td>
<td>1, 32</td>
<td>.261</td>
</tr>
<tr>
<td>Base Rate x Feedback 1</td>
<td>9.1***</td>
<td>2, 24</td>
<td>.02</td>
</tr>
<tr>
<td>Base Rate x Feedback 2</td>
<td>20.6***</td>
<td>1, 32</td>
<td>.261</td>
</tr>
<tr>
<td><strong>FA repeated measures for block means</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blocks 1</td>
<td>6.6***</td>
<td>6, 144</td>
<td>.007</td>
</tr>
<tr>
<td>Blocks 2</td>
<td>4.76***</td>
<td>14, 448</td>
<td>.018</td>
</tr>
<tr>
<td>Block x Base Rate 1</td>
<td>6.8***</td>
<td>6, 144</td>
<td>.007</td>
</tr>
<tr>
<td>Block x Base Rate 2</td>
<td>2.98***</td>
<td>14, 448</td>
<td>.018</td>
</tr>
<tr>
<td>Blocks x Feedback 1</td>
<td>2.34**</td>
<td>14, 448</td>
<td>.018</td>
</tr>
<tr>
<td><strong>FA block means for feedback procedure only</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blocks 1</td>
<td>3.0***</td>
<td>6, 120</td>
<td>.008</td>
</tr>
<tr>
<td>Blocks 2</td>
<td>3.22***</td>
<td>14, 224</td>
<td>.015</td>
</tr>
<tr>
<td>Blocks x Base Rate 1</td>
<td>3.2*</td>
<td>6, 120</td>
<td>.008</td>
</tr>
<tr>
<td>Blocks x Base Rate 2</td>
<td>2.22**</td>
<td>14, 224</td>
<td>.015</td>
</tr>
</tbody>
</table>

*(table continues)*
Appendix B (continued)

<table>
<thead>
<tr>
<th>Condition, variable, and experiment</th>
<th>F</th>
<th>df</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Letter-digit (continued)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blocks 2</td>
<td>3.8***</td>
<td>14,224</td>
<td>.021</td>
</tr>
<tr>
<td>Correct recognitions block means within feedback procedure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base rate 1</td>
<td>7.8*</td>
<td>1, 8</td>
<td>.11</td>
</tr>
<tr>
<td>Blocks 1</td>
<td>3.1**</td>
<td>6, 48</td>
<td>.04</td>
</tr>
<tr>
<td>Correct recognitions block means within control procedure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blocks 1</td>
<td>4.9***</td>
<td>6, 120</td>
<td>.02</td>
</tr>
</tbody>
</table>

Words

<table>
<thead>
<tr>
<th>Source (new, study, series)</th>
<th>F</th>
<th>df</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>119.4***</td>
<td>2, 16</td>
<td>.04</td>
</tr>
<tr>
<td>2</td>
<td>511.6***</td>
<td>2, 16</td>
<td>.004</td>
</tr>
<tr>
<td>Source x Feedback</td>
<td>11.0***</td>
<td>2, 16</td>
<td>.004</td>
</tr>
<tr>
<td>FA block means</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blocks 2</td>
<td>4.97***</td>
<td>15, 120</td>
<td>.01</td>
</tr>
<tr>
<td>Within feedback procedure, blocks</td>
<td>3.0***</td>
<td>15, 60</td>
<td>.007</td>
</tr>
<tr>
<td>Within control procedure, blocks</td>
<td>3.4</td>
<td>15, 60</td>
<td>.017</td>
</tr>
</tbody>
</table>

Note. Order of stimuli, included as a blocking factor in the analyses of variance, yielded no significant main effects or interactions with other factors in either experiment. This finding does not imply that order is unimportant—it only implies that when orders are randomly drawn and trial sequences are relatively long, conclusions about effects of experimental variables on condition means are not limited to the particular orders used. The degrees of freedom and F values summarized here are for effects and trends denoted as significant or marginally significant in the text. FA = false alarm.

*p < .05. **p < .01. ***p < .001.

Appendix C

Augmentations of the Array Model for Feedback Conditions

Version 1: Bias Adjustment

Of the alternative augmentations of the array model we have considered, perhaps the simplest is to add the assumption that a learner's criterion, or bias, for calling test stimuli “old” is modified by feedback. To be specific, we assume that during a trial series in which the probability of feedback denoting an old item has some value \( \pi \), the value of \( b \) (defined by Equation 6) changes in accord with the function

\[
b_{n+1} = b_n + \theta(\pi - b_n),
\]

where the subscript \( n \) has been added to index the trial number, and the “learning parameter,” \( \theta \), is a constant with a value between 0 and 1. It is apparent that, on the average, the difference between the current value of \( b \) and the old stimulus base rate, \( \pi \), is reduced by a fraction \( \theta \) on each trial, so that, over a series of trials with the feedback procedure, the bias measure, \( b \), will approach the old stimulus base rate (i.e., \( b \) will approach \( \pi \)). As the learning process represented by Equation C1 becomes asymptotic, the probability of a false alarm does not approach the base rate but, denoting false-alarm probability by \( P \) for brevity, it is straightforward to show that a linear relation obtains between \( (1 - P)/P \) and \((1 - \pi)/\pi \), that is,

\[
(1 - P)/P = [(1 - \pi)/\pi] s_0/Ns,
\]

where \( N \) denotes the total number of item representations stored in the old memory array and \( s \) is average interitem similarity. As \( N \) becomes large, the right-hand side of Equation C2 approaches zero, and therefore \( P \) approaches unity.

Version 2: Contextual Learning

In this augmentation, it is assumed that recognition performance depends on a mixture of outputs from the process represented in the array model and a contextual learning process. The form of the mixture model can be summarized as follows. The probability of making an old response if the learner attends only to the background context and not to the item presented on a trial \( n \) of a series will be denoted \( P_{n} \), and for a trial series in which the probability of feedback denoting an old item has some value \( \pi \), the value of \( P_{n} \) is assumed to change in accord with the function

\[
P_{n+1} = P_{n} + \theta(\pi - P_{n}),
\]

which dates back to the earliest applications of stimulus sampling theory to probability learning (Estes & Straughan, 1954, p. 226, Equation 3). The learning parameter, \( \theta \), which has a value between 0 and 1, is the only free parameter that we added to the three parameters of the array model in order to interpret the feedback.
condition. Equation C3 has the same form as Equation C1 and implies that over a test series under feedback, the value of \( p \), the probability of a participants's making an old response when attending solely to the background context, will approach \( \pi \), the old stimulus base rate.

We incorporate the contextual learning process into the array model by assuming that on each trial of a recognition test series in the feedback condition, the participant attends both to the item presented and to the background context. If we denote the total similarity of a test item to the old memory array by \( S_i \) and the probability of an "old" response to the item in the array-model process by \( P_i \), where

\[
P_i = \frac{S_i}{(S_i + S_0)},
\]  
(C4)

then the unconditional probability of an "old" response, \( P(\text{Old}) \), on the trial is assumed to be given by the expression

\[
P(\text{Old}) = P_i + (1 - P_i)p
\]
\[
= p + (1 - p)P_i.
\]  
(C5)

To derive the asymptotic probability of an "old" response to a new stimulus, we need only substitute for \( p \) in Equation C5 its asymptotic value, \( \pi \), and for \( P_i \) its equivalent, \( Ns/(Ns + s_0) \), from Equation 6, obtaining

\[
P(\text{Old}) = \pi + (1 - \pi)Ns/(Ns + s_0).
\]  
(C6)

The quantity \( Ns/(Ns + s_0) \) approaches unity as \( N \) becomes large, so the righthand side of Equation C6 approaches \( \pi + (1 - \pi) = 1 \). Thus, false-alarm probability is not predicted to approach the base rate, \( \pi \), asymptotically, but, just as in the bias model, it is predicted to approach unity.

Whether this prediction about asymptotes can be tested experimentally is not clear. The difficulty is that, for all of the stimulus types we studied, at least, the rate of memory decay, as indexed by the parameter \( \alpha \) of the array model, is great enough that the number of stimulus representations stored in the memory array, \( N \), grows over trials at a slower rate than the number of stimuli presented and, even over a much longer trial series than is ever seen in experimental studies, may not become large enough for \( P(\text{Old}) \) to approach asymptote.

The result of fitting the augmented models, in both their \( \delta_{\text{learn}} \) and \( \delta_{\text{pattern}} \) versions, to our feedback data is summarized in Table C1 in terms of mean absolute deviations of predicted from observed response percentages, averaged over Experiments 1 and 2. The advantage for the context learning over the bias-adjustment model looks quite convincing, but no formal statistical test of the difference in goodness of fit is available, so the result should be taken as provisional, pending opportunities for replication.

Received November 3, 1993
Revision received September 1, 1994
Accepted September 14, 1994

<table>
<thead>
<tr>
<th>Model</th>
<th>( \delta_{\text{learn}} )</th>
<th>( \delta_{\text{pattern}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>New</td>
<td>Study</td>
</tr>
<tr>
<td>Context</td>
<td>1.50</td>
<td>1.50</td>
</tr>
<tr>
<td>Bias</td>
<td>3.25</td>
<td>8.50</td>
</tr>
</tbody>
</table>

Note. Data are percentages of "old" responses by stimulus source, averaged over Experiments 1 and 2.