Journal of Experimental Psychology:
Human Perception and Performance
1982, Vol. 8, No. 6, 834-864

0096-1523/82/0806-0834$00.75

Experimental Test of Contemporary Mathematical Models of Visual Letter Recognition

James T. Townsend
Purdue University

F. Gregory Ashby
Ohio State University

A letter confusion experiment that used brief durations manipulated payoffs across the four stimulus letters, which were composed of line segments equal in length. The observers were required to report the features they perceived as well as to give a letter response. The early feature-sampling process is separated from the later letter-decision process in the substantive feature models, and predictions are thus obtained for the frequencies of feature report as well as letter report. Four substantive visual feature-processing models are developed and tested against one another and against three models of a more descriptive nature. The substantive models predict the decisional letter report phase much better than they do the feature-sampling phase, but the best overall 4×4 letter confusion matrix fits are obtained with one of the descriptive models, the similarity choice model. The present and other recent results suggest that the assumption that features are sampled in a stochastically independent manner may not be generally valid. The traditional high-threshold conceptualization of feature sampling is also falsified by the frequent reporting by observers of features not contained in the stimulus letter.

Knowledge of how people identify letters in their native alphabet is of fundamental interest, not only in being a special important case of more general pattern perception behavior but also because some amount of letter recognition must play a role in reading behavior.

The present study focuses on the recognition and confusion of single letters constructed from line segments of equal length and presented at brief durations of display. An experimental paradigm is used that requires a report of perceived features as well as of the letter believed to have been present. This paradigm permits a tentative decomposition of the overall perceptual process into an earlier sensory phase and a subsequent decision phase. A number of important hypotheses about the recognition process, most of them embedded in mathematical models, are tested in this context.

Feature models have often been the preferred instrument for providing a mathematically specified processing explanation of recognition–confusion behavior (e.g, Geyer & DeWald, 1973; Townsend, Hu, & Ashby, 1980; Wandmacher, 1976; Gibson, Osser, Schiff, & Smith, Note 1; Rumelhart, Note 2). This preference is probably due in part to the ability of the models to provide a reasonably flexible structure while still remaining sufficiently parsimonious so that their parameters can be estimated. With regard to the conception of features as lines, combinations of lines, and similarly elementary geometric aspects of letters, the seminal physiological work of Hubel and Wiesel (1962) has been singularly influential.

Models of a more descriptive, but less substantively detailed nature have also been useful in the study of recognition and confusion (e.g., Broadbent, 1967; Pachella, Smith, & Stanovich, 1978; Townsend, 1971a, 1971b). In the present work, a broad class of feature-processing models are tested in addition to

Some of the early analyses of this study were reported at Midwestern Psychological Association, Chicago, 1976. Later stages of the investigation were supported by National Institute of Health Grant 1R03MH28551-01 and by National Science Foundation Grant 7920298, and much computer time was provided through the Purdue Research Foundation. Final typing took place while the first author was a visiting scholar at University of California at Irvine, 1982.

We are indebted to Stephen Williams for aid in carrying out the experiment.

Requests for reprints should be sent to James T. Townsend, Department of Psychological Sciences, Purdue University, West Lafayette, Indiana 47907.

a well-known group of descriptive models. Remarks concerning possible representation of hierarchical mechanisms and gestalt properties are made later.

It is our belief that recognition behavior is best investigated through the use of mathematical models. This manner of investigation has the advantages of: (a) permitting the rigorous testing of important psychological concepts embedded in a full model of behavior in appropriate tasks, (b) providing for the testing of distinct full theoretical models against one another, and (c) explicating structure of data through examination of patterns of parameter estimates and other model analyses. Moreover, the testing of two or more models against one another avoids the pernicious statistical and philosophical difficulties associated with attempting to verify or falsify a single model or theory. Where possible, we feel models that make predictions in closed mathematical form are preferable to models that require computer simulation. All the models below are of this ilk.

The present report is self-contained, but general background material in theoretical issues in this area can be found in Garner (1978), Massaro and Schmuller (1975), and Reed (1973). Townsend and Landon (in press) survey the most used mathematical models of recognition-confusion behavior. The next section gives a brief introduction to complete identification experiments and confusion matrices and outlines the classes of models investigated in the present report.

Letter Recognition: Complete-Identification Experiments and Mathematical Models

In a complete-identification experiment, there is a unique response associated with each distinct stimulus rather than a rougher categorization of the stimuli where several stimuli may have the same response. Perception is degraded in some way, often by very brief exposure, and performance is expressed in terms of a confusion matrix. If there are N stimuli and N responses (e.g., $N = 26$ for the English alphabet), then a confusion matrix has N rows, one for each stimulus, and N columns, one for each response. These are placed in such a way that the diagonals are used to record correct responses and the off-diagonals are used to record incorrect responses, or confusions.

When the entries in each row are divided by the sum across columns in that row, the result is a proportion in each cell that gives $\hat{P}(R_j|S_i) = \hat{c}_{ij}$, the proportion of times that presentation of stimulus S_i resulted in response R_j, i and j running from 1 to N. Of course, c_{ii} is the proportion correct for stimulus S_i. These proportions are estimates of the true underlying conditional probabilities, $P(R_j|S_i) = c_{ij}$. Formal models may be used to predict the c_{ij} and tested against their estimates $\hat{c}_{ij}$.

As Fisher, Monty, and Glucksberg (1969) have pointed out, patterns of confusions in such matrices may differ widely depending on particular experimental conditions. Likewise, Garner and Haun (1978) have shown how different types of confusion errors can result with the same letters or stimulus patterns but under different presentation conditions. The use of mathematical models may sometimes be of help in instances such as these. For example, it was found in an earlier study that much of the confusion variability obtained with the same alphabet but under differing masking conditions or with different observers could be attributed to variation in response bias rather than variation in sensory factors (Townsend, 1971a, 1971b).

We are concerned here with the situation where the letters have been learned, and it is therefore clear that any model of the recognition process must have three major aspects. The first is that some type of stimulation or input must occur. The second is that information about the set of letters to be identified must be preserved in some form within the perceiving system. Third, at least part of the information present in the stimulus must make contact with, or be matched against, part of the stored information. So-called "passive" recognition theories use the matching process heavily, whereas "analysis-by-synthesis" theories use partial matching results along with construction roles to synthesize potential answers. The latter have received little mathematical treatment or application to confusion experiments, so are excluded from further discussion here.

All the models in this article can be char-

a well-known group of descriptive models. Remarks concerning possible representation of hierarchical mechanisms and gestalt properties are made later.

It is our belief that recognition behavior is best investigated through the use of mathematical models. This manner of investigation has the advantages of: (a) permitting the rigorous testing of important psychological concepts embedded in a full model of behavior in appropriate tasks, (b) providing for the testing of distinct full theoretical models against one another, and (c) explicating structure of data through examination of patterns of parameter estimates and other model analyses. Moreover, the testing of two or more models against one another avoids the pernicious statistical and philosophical difficulties associated with attempting to verify or falsify a single model or theory. Where possible, we feel models that make predictions in closed mathematical form are preferable to models that require computer simulation. All the models below are of this ilk.

The present report is self-contained, but general background material in theoretical issues in this area can be found in Garner (1978), Massaro and Schmuller (1975), and Reed (1973). Townsend and Landon (in press) survey the most used mathematical models of recognition-confusion behavior. The next section gives a brief introduction to complete identification experiments and confusion matrices and outlines the classes of models investigated in the present report.

Letter Recognition: Complete-Identification Experiments and Mathematical Models

In a complete-identification experiment, there is a unique response associated with each distinct stimulus rather than a rougher categorization of the stimuli where several stimuli may have the same response. Perception is degraded in some way, often by very brief exposure, and performance is expressed in terms of a confusion matrix. If there are N stimuli and N responses (e.g., $N = 26$ for the English alphabet), then a confusion matrix has N rows, one for each stimulus, and N columns, one for each response. These are placed in such a way that the diagonals are used to record correct responses and the off-diagonals are used to record incorrect responses, or confusions.

When the entries in each row are divided by the sum across columns in that row, the result is a proportion in each cell that gives $\hat{P}(R_j|S_i) = \hat{c}_{ij}$, the proportion of times that presentation of stimulus S_i resulted in response R_j, i and j running from 1 to N. Of course, c_{ii} is the proportion correct for stimulus S_i. These proportions are estimates of the true underlying conditional probabilities, $P(R_j|S_i) = c_{ij}$. Formal models may be used to predict the c_{ij} and tested against their estimates $\hat{c}_{ij}$.

As Fisher, Monty, and Glucksberg (1969) have pointed out, patterns of confusions in such matrices may differ widely depending on particular experimental conditions. Likewise, Garner and Haun (1978) have shown how different types of confusion errors can result with the same letters or stimulus patterns but under different presentation conditions. The use of mathematical models may sometimes be of help in instances such as these. For example, it was found in an earlier study that much of the confusion variability obtained with the same alphabet but under differing masking conditions or with different observers could be attributed to variation in response bias rather than variation in sensory factors (Townsend, 1971a, 1971b).

We are concerned here with the situation where the letters have been learned, and it is therefore clear that any model of the recognition process must have three major aspects. The first is that some type of stimulation or input must occur. The second is that information about the set of letters to be identified must be preserved in some form within the perceiving system. Third, at least part of the information present in the stimulus must make contact with, or be matched against, part of the stored information. So-called "passive" recognition theories use the matching process heavily, whereas "analysis-by-synthesis" theories use partial matching results along with construction roles to synthesize potential answers. The latter have received little mathematical treatment or application to confusion experiments, so are excluded from further discussion here.

All the models in this article can be char-

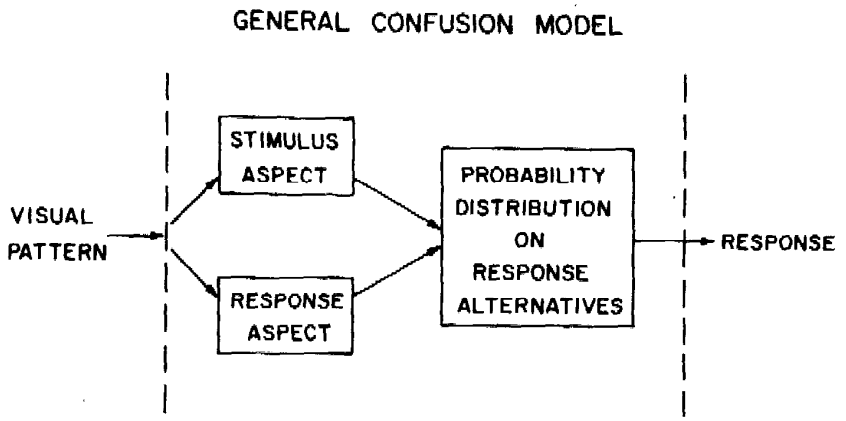

Figure 1. Schematic of a general confusion model.

acterized as special cases of a general confusion model as shown in Figure 1. The broadest property of the model is its supposition of two major aspects of pattern recognition, one related to attributes of the stimuli and the other to attributes associated with the responses. The first is often referred to as the sensory structure of the model and typically involves parameters which should be affected primarily by stimulus variables such as intensity, clarity, degree of masking, and stimulus similarity. The second reflects response-related variables such as frequency, motivational variables (e.g., payoff structure), and other aspects of the situation related to response preference or decision bias.

Although the term "sensory" may be reasonably accurate in simple psychophysical situations (e.g., in simple detection experiments; Green & Swets, 1966), the attendant connotations of extreme peripherality and simplicity may be misleading in more complex recognition experiments. Therefore, the less loaded term "stimulus" will usually be used in the present study. In all models but the choice model, the stimulus and response aspects take the form of separate subprocesses, and it may be that similar interpretations are possible for the choice model.

Figure 2 shows how the various models of this study can be "derived" as special cases of the general confusion model. The first two major subdivisions shown are the *descriptive* and the *substantive* models.

The descriptive models are general devices that can be used to test or reflect broad principles within a typically highly parametered framework, whereas the present substantive models attempt to make more specific assumptions about feature-processing mechanisms. The latter can be used to provide testable predictions emanating from the specific assumptions. The descriptive models on the other hand provide large-scale qualitative information about confusions and also yield quantitative data concerning sensory and response-bias tendencies from their parameters.

There were several purposes involved in including the descriptive models. The most important was to provide a set of references with which to compare the performance of the substantive feature models. A secondary purpose was to evaluate the overall accuracy of prediction of the descriptive models. Third, it was of interest to compare the performance of the choice model (Luce, 1963) with the overlap model under suitable experimental conditions, which is discussed below, and to compare both the choice model and the overlap model with the all-or-none model (Townsend, 1971a, 1971b).

The second branch on the left in Figure 2 splits into process versus nonprocess models. Process model is a fuzzy concept that is used a great deal but is difficult to pin down. Here "process model" means that the model specifies some hypothetical temporal order of operations in the recognition task. Thus, the choice model is tentatively classified as a nonprocess model because no satisfactory temporal formulation was originally given. Possibilities toward this end have been put forth elsewhere (Townsend, Note 3). The so-

phisticated guessing model supposes that presentation of a stimulus results in partial information that is compatible with one or more stimulus alternatives, the latter alternatives being called the confusion set. The response is then selected from the confusion set by guessing from among the alternatives, perhaps with unequal probabilities (see, e.g., Broadbent, 1967). It is also typically assumed that the correct response is always contained in the confusion set. The all-or-none and overlap models may be designated as models of this type.

On the right side of Figure 2, one can see that the present substantive models are viewed as process models because the detailed assumptions are about the mechanisms of processing. All the substantive models considered here are feature-processing models that are defined in detail below.

The divisions of Figure 2 should not be thought of as absolute. For example, some feature-processing models can be viewed as special cases of sophisticated guessing models (thus, the dotted connecting line in Figure 2). Also, sophisticated guessing models can sometimes be given feature interpretations. The feature-distance models, the multicomponent model, and the hit-ratio models, below, all presume that feature information leads to the establishment of a confusion set from which the final choice is made according to a set of bias probabilities. Further, both the multicomponent model (Rumelhart, Note 2) and the hit-ratio model (Geyer & DeWald, 1973), in their original formulation, make assumptions equivalent to presuming that the correct alternative is always included in the confusion set. However, for present purposes it is convenient to separate the sophisticated guessing models from the substantive models. Below, it is also apparent that circumstances may arise where it is psychologically reasonable to assume that confusion sets may exist that do not contain the presented stimulus. The latter possibility violates the usual formulation of the sophisticated guessing model.

The stimulus- versus response-related aspects assumption of the present process models is strong in that, given a particular perceptual state on a given trial, the stimulus has no further influence on the selected response. This assumption amounts to a Markov axiom with regard to the early and later phases of processing.

In the substantive feature models, the stimulus-related process is called the "feature-sampling process," and the response-related process is called the "decision process." The general order of events occurring within systems described by the present substantive models is given in more detail below.

Current feature models tend to make similar and strong assumptions about the feature sampling. They differ mainly in their assumptions about what happens after a set of features is sampled from the stimulus pat-

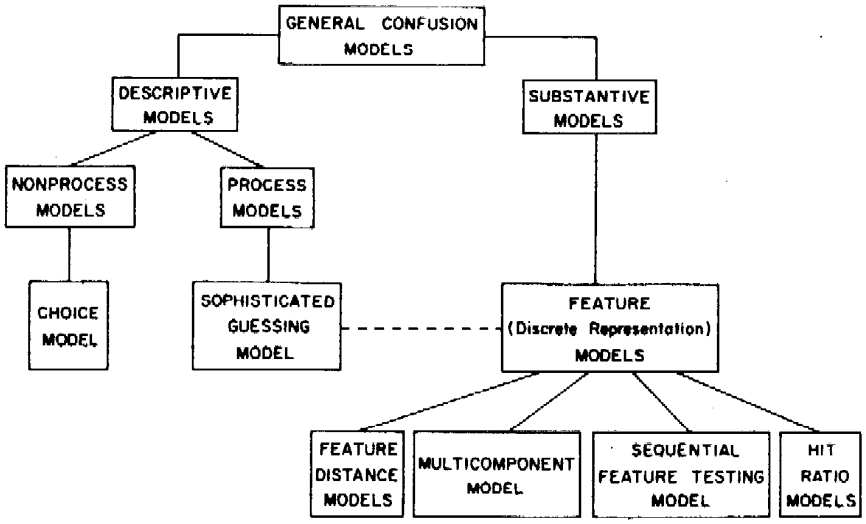

Figure 2. Relations between various well-specified models of confusion.

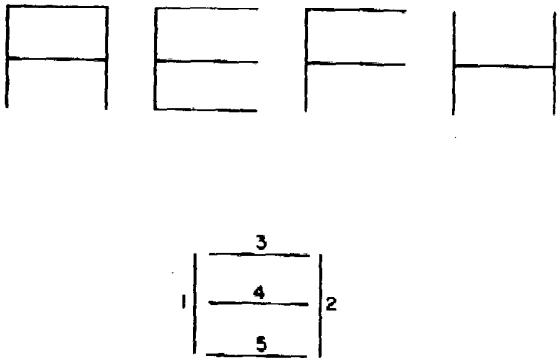

Figure 3. The four-letter stimuli used and the five line features of equal length from which the stimuli were constructed.

tern. For this reason as well as reasons connected with testing assumptions in the context of the present experiment, it is convenient to separate the presentation and tests of the feature sampling and decision process in later sections. The next section introduces the experiment in the context of the present theoretical questions. The experiment was designed to test a large class of recognition models, including the stimulus versus response aspects, the two-phase Markov representation of these, and most of the models shown at the terminal points of Figure 2. However, only certain special cases of the sophisticated guessing model are tested.

Structure and Purposes of the Experiment

The basic experiment consisted of complete identification of a small set of stimulus letters constructed from a set of line segments equal in length. These were the letters A, E, F, and H, which are shown along with the elemental feature population in Figure 3. As noted earlier, a novel aspect of the study was the inclusion of a listing by observers of the features they believed they saw on each trial, before their letter report.

As most previous model applications have been on the overall confusion matrices, it has been difficult to evaluate the classic separate assumptions about the feature-sampling process and the decision process. The present experiment requires the report of the features by observers as well as of the letter, and therefore it represents an attempt to expose more detail of the underlying processing in the context of a typical recognition paradigm. With this technique, it is also possible to examine the Markov assumption, made by all extant detailed feature-recognition models, that the feature sample carries all the information from which a letter decision is made. Thus if observers' feature reports accurately represent the feature samples and if the latter are responsible for the subsequent decision, then the response, when conditionalized on the feature report, should be independent of the original stimulus letter.

In addition, previous mathematical models of the feature-sampling process were "high threshold" in the sense of assuming that features not present in the stimulus are never present in the feature sample. The feature reports allow an opportunity to test this assumption directly, and our feature-sampling models permit estimation of such "ghost feature" sampling probabilities. The feature report aspect of the procedure may be thought controversial. For example, perhaps an observer "recognizes" in the form of a letter then decides what features to report. First, the results from a control condition do not support this view. Second, the conclusions based on the feature reports are compatible with a number of other recent investigations as is discussed in the section on feature perception.

Another important characteristic of the design was manipulating the motivational conditions by altering the payoff structure. There have been no experiments to our knowledge manipulating bias or sensory factors in order to test whether model parameters exhibit the predicted types of invariance or change. The only exception appears to be a study in our laboratory that varied the size of the stimulus alphabet—a variation that primarily affected sensory structure (Townsend & Landon, 1982). Thus, if only motivational conditions are experimentally manipulated, the true model should reveal alterations only in its response-bias parameters, with its stimulus-aspect parameters remaining constant. In contrast, an untrue model should exhibit change in both its stimulus-aspect as well as its response-aspect (bias) parameters. This type of manipulation is clearly important in providing a more rigorous test to models than is afforded by fits

to a single confusion matrix. The present study includes experimental variation of response biases via two different payoff situations.

Previous work has emphasized tests within the context of a single substantive feature model, although these have sometimes been compared with descriptive models (e.g., Geyer & DeWald, 1973; Rumelhart, Note 2), and occasionally variations within a single model have been analyzed (e.g., Wandmacher, 1976). This study attempts to test a wide variety of models of the feature-sampling process as well as of the decision process. The following section describes the models under investigation in more detail.

Theoretical Developments

In this section, we begin by introducing the descriptive models. This introduction is followed by a description of the various substantive models of the sensory feature-sampling process leading from the stimulus letter to a well-defined perceptual or "sensory" state of perceived features. Last, the different substantive feature models of the decision process, leading from the sampled feature state via a confusion set of response alternatives to a letter response, is presented.

Descriptive Models

Townsend (1971a, 1971b) has applied three descriptive models to alphabetic confusion matrices, the all-or-none model, the overlap model, and Luce's (1963) choice model. As noted above, these models may be applied with fewer detailed psychological assumptions than may the feature models. Even so, all three use stimulus, or sensory, parameters and a mathematically independent set of decision, or bias, parameters. Thus, in all cases, response biases are independent of the stimulus presented.

Their flexibility, which is due to their generality and a relatively large number of free parameters, permits investigation of broader issues and various psychological influences. For example, it will be seen that the all-or-none model can be viewed as a simple type of template model and hence can be used to test this notion. Both Townsend (1971a, 1971b) and Rumelhart (Note 2) have found the choice model to provide the best fits, although the overlap model was close in both instances. Indeed, Holbrook (1975) has used the Townsend (1971a, 1971b) similarity estimates from the choice model as a yardstick with which the performance of more specific models is compared.

All-or-None Model

The all-or-none model (Townsend, 1971b), sometimes referred to as the threshold model, the pure perceptibility model (Smith, Note 4), and the pure guessing model (Broadbent, 1967), assumes that following presentation of a stimulus S_i, the observer with probability P_i recognizes the stimulus perfectly. With probability $1 - P_i$, the observer is thrown into a zero-information state but still may correctly guess the proper response with probability of success h_i. The confusion matrix is then given by:

$$c_{ij} = \begin{cases} (1 - P_i)h_j, & i \neq j \\ P_i + (1 - P_i)h_i, & i = j, \end{cases} \quad (1)$$

where

$$0 \leq P_i \leq 1,\ 0 \leq h_j \leq 1, \text{ and } \sum_{j=1}^{N} h_j = 1.$$

The all-or-nothing perceptual property of this model may be given a strict, simple template interpretation, or it could be derived from other processing situations. The template version is clear: A perfect match of the stimulus occurs, or no information at all is transmitted.

In one alternative interpretation, partial information (perhaps a subset of features) is processed on some trials, but on every trial there is either sufficient information to denote the correct response with certainty, or the present information is totally unhelpful. For instance, if each stimulus letter contained a single subset of unique features plus other noninformative features, P_i would equal the probability of extracting one or more of the unique features from stimulus i. Another possibility is where partial information is not used by the observer for some reason.

The all-or-none model requires estimation of $2N - 1$ free parameters, or seven for the

present experiment, versus $N(N-1)$, or 12 degrees of freedom in the data.

It may be observed that the all-or-none model is one type of confusion model wherein off-diagonal cells of the confusion matrix are assumed to be decomposable into a product of a stimulus factor multiplied by a response factor: $c_{ij} = a_i \times b_j$. These have undergone investigation recently by Falmagne (1972), Townsend (1978), and Wandmacher (1977).

Overlap Model

The overlap model assumes that either perfect information or two-way partial information is acquired by the observer on each trial (Townsend, 1971b). With probability E_{ii}, stimulus S_i is recognized perfectly and elicits the correct response R_i. With probability E_{ij}, the observer is in a pairwise confusion state with uncertainty as to whether the stimulus was S_i or S_j. When in this state of two-way confusion, the observer is assumed to guess stimulus S_j according to the ratio of the response biases involved, $g_j/(g_i + g_j)$. The confusion matrix is given by:

$$c_{ij} = \begin{cases} E_{ij}\left(\dfrac{g_j}{g_i + g_j}\right), & i \neq j \\ E_{ii} + \sum\limits_{\substack{k \neq i \\ k=1}}^{N} E_{ik}\left(\dfrac{g_i}{g_i + g_k}\right), & i = j, \end{cases} \tag{2}$$

where $0 \leq E_{ij} \leq 1$, $0 \leq g_j \leq 1$, $\sum\limits_{i=1}^{N} g_i = 1$, $\sum\limits_{j=1}^{N} E_{ij} = 1$, for $i = 1, N$ and $E_{ij} = E_{ji}$. The overlap model contains $[N(N+1)]1/2 - 1$ free parameters to be estimated, or nine here.

The overlap model has been shown to be falsifiable in principle with respect to the choice model when moderately to highly confusable stimuli are used. (Pachella et al., 1978; Townsend, 1971b). One of the rather tight constraints imposed on confusion data by the overlap model is that

$$c_{jj} \geq \sum_{\substack{i \neq j \\ i=1}}^{N} c_{ij} \tag{3}$$

for $1 \leq j \leq N$; that is, the column sum of off-diagonal entries must be less than or equal to the diagonal (proportion correct) entry. The choice model is not constrained by this inequality. Heretofore, however, the overlap and choice models have been quite close in their ability to predict empirical confusion matrices, and the reason may be that the stimuli were not sufficiently confusable relative to the luminance and duration of the display to test the two. As is later seen, the letters used in the present experiments were highly confusable and should provide a firm test between the choice and overlap models.

Townsend (1971b) suggested a generalized overlap model that combined the pairwise confusion state of the overlap model with the no-information state of the all-or-none model, plus the perfect information state included in both. Pachella et al. (1978) developed a model of this type that they applied to a speed–accuracy confusion experiment. Generalized overlap models were not considered here although we believe them to be of value in certain contexts. One may imagine rough feature interpretations of the overlap model, in which features are shared by no more than two stimuli.

Similarity-Choice Model

This model was proposed by Luce (1963); like the above models, it contains sensory-(or stimulus-)similarity parameters and response-bias parameters. It has sometimes been referred to as the biased-choice model (e.g., Pachella et al., 1978). However, because several versions of the choice model contain bias structure, but not all include similarity structure, we prefer the present name. This model is rather different in conception from the simpler recognition model proposed earlier by Luce (1959). The primary likeness of the two models is that both the sensory and bias parameters are assumed to be "strengths" lying on a ratio scale.

The similarity-choice model proposes that given a presentation of a stimulus S_i, the strength of response R_j is given by a product of the sensory parameter M_{ij}, which represents the similarity between stimulus S_i and S_j, and the response bias b_j, which is independent of stimulus presentation. The response probability of R_j given S_i is denoted

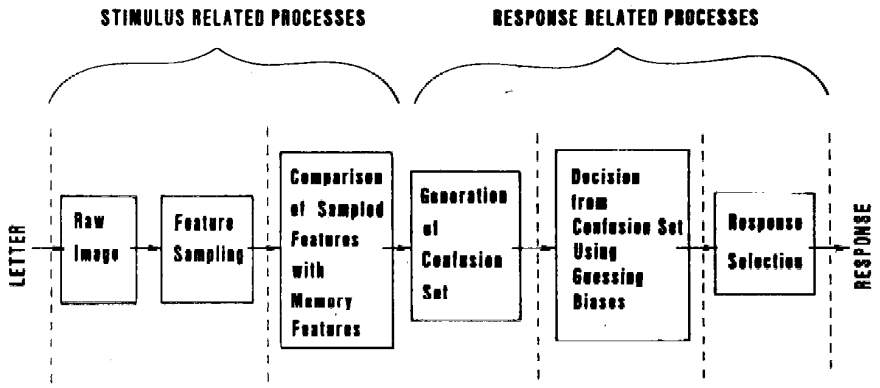

Figure 4. Schematic of the letter-recognition process as evidenced by the substantive feature-processing models.

by the ratio of the scale value of R_j, namely $M_{ij}b_j$, to all other possible response strengths. Therefore the confusion matrix becomes:

$$c_{ij} = \frac{M_{ij}b_j}{\sum_{k=1}^{N} M_{ik}b_k} \quad \text{for} \quad 1 \leq i, j \leq N, \quad (4)$$

where $0 \leq M_{ij} \leq 1$, $0 \leq b_j \leq 1$, and $M_{ii} = 1$ for $i = 1, N$, and $M_{ij} = M_{ji}$. The choice model contains the same number of free parameters as does the overlap model, namely, $[N(N + 1)]/2 - 1$. This property facilitates comparison of the two models. As in the overlap model, the similarity parameter is symmetric, $M_{ij} = M_{ji}$. Finally, note that the similarity of a stimulus S_i with itself equals 1 (i.e., $M_{ii} = 1$). The choice model is further discussed later.

Lappin (1978) has argued cogently for expanding the role of the bias parameters to include response pairwise relation in this and certain other models. A detailed examination of the "pro" and "con" arguments of such a move is outside the scope of the present work. However, use of the traditional model permits comparison with previous work as well as with the equal-parametered overlap model. More important, it is seen below that the traditional model, with its smaller number of parameters, continues to perform acceptably.

The next developments deal with the substantive feature models. As mentioned above, we separate the exposition of models of the feature-sampling process from the decisional processes.

Substantive Feature-Processing Models

Although most models of detection and recognition posit a stimulus-related (sensory) and a response-related (decision) process, as does the general confusion model, it is likely that the representation may be too simplistic a description of the actual functioning of the human perceptual system. The substantive feature-processing models are a step in the right direction because they break down the overall system into a more detailed set of subsystems and mechanisms. Figure 4 illustrates this breakdown. The vertical, dotted lines indicate the analysis used to present and discuss the processing mechanisms, whereas the brackets indicate the stimulus- versus response-related distinction.

Models of the Feature-Sampling Process

The current models (or submodels) of the feature-sampling process make quite strong assumptions that we examine below. The part of the experiment in which observers reported the features they believed to be present in the stimulus presentation is an attempt to achieve better observability into the feature-sampling process than is had with the usual procedure of simply requiring a letter report. It is true that in order to make a report of the "seen" features, the sensory information must travel through some decisional process as to whether or not to report any given feature. However, it is also true that there may exist decisional influences on the sampled feature set in the ordinary case where only the letter is reported. No current

models of the feature-sampling process that have been fit to confusion matrices explicitly allow for decisional or motivational influences, although the assumption by Rumelhart (Note 2) that a functional feature is accepted if a single component of that feature is extracted could easily be relaxed into a criterial model of feature sampling. The feature reports can also be viewed on a more phenomenological level as simply providing data about what aspects of the letters are seen but are typically lost by the necessity of categorizing the perceptual information into the various letter responses.

We shall describe and analyze the feature-sampling models and the data on feature reports from the same naive point of view taken by the models themselves and then consider proper cautions as well as possible implications for future models.

Almost all extant feature-sampling models are simple variations based on assumptions of independence and invariance of various types. These assumptions are, at present, of necessity, because data yielding a simple view of sensory perceptual processes uncluttered by decision processes are scant. Most existing models assume that the extraction or sampling of one feature from the stimulus is probabilistically independent of extraction of others. We call this, the most critical of the present sampling assumptions, *sampling independence*, and one of the aims of the present study is to test it.

In all, five sensory models were tested. For all models, let alpha equal the probability of sampling a feature that exists in the stimulus; with probability $1 - \alpha$, the feature is *lost*. Let gamma equal the probability of sampling a feature that does not exist in the stimulus. That is, gamma represents the likelihood that one of the features from which the stimuli are constructed arises from noise in the display or sensory mechanisms rather than from the current stimulus itself. When this event occurs, the "extracted" feature is called a *ghost* feature. Note that when gamma is greater than 0, we have what amounts to a low-threshold model of feature detection. There are no other examples of low-threshold or ghost-feature models in the literature.

All present models postulate that the probability of sampling, say, feature j is constant across stimulus letters in which it appears. That is, there are no gestalt or capacity effects, as when one letter contains more features than another, imposed by the separate unique letters. This assumption will be referred to as *across-letter invariance*.

A third assumption premises that the feature-sampling probabilities are all equal for the various features. This constraint is assumed in the first and second feature-sampling models listed below for both natural and ghost features and in the fourth model on ghost, but not natural features. Thus, this assumption is testable by examination of alpha and gamma parameter estimates. We call this assumption *across-feature invariance*.

The models, which were fit to data from the present experiment, are defined as follows:

1. Alpha is constant; gamma equals 0. Here it is assumed that all features have equal probability of being sampled and, further, that ghost features never appear. This model is proposed by Rumelhart (Note 2; for the special case where all features are of equal length as they are in the present experiment).

2. Alpha and gamma are constant. Now it is assumed all ghost features (features not contained in the presented stimulus) have an equal, nonzero probability of being sampled.

3. Alpha can vary; gamma equals 0. Each of the five features has an unique probability of being sampled when it is actually present, but ghost features cannot show up in the feature sample. This model contains five free parameters with our stimuli and corresponds to Geyer and DeWald's (1973) most general sensory assumptions.

4. Alpha can vary; gamma is constant. Here we have six free parameters.

5. Alpha and gamma can vary. This case adds only two free parameters, as two of the five features exist in all stimuli and therefore can never be ghost features.

In addition to the feature-sampling model fits, probabilistic independence and the two kinds of invariance can be evaluated directly by chi-square tests in our data on feature report.

There have been few attempts to directly

test sampling independence. We discuss what literature exists in the context of our results below.

Substantive Models of the Decision Process

Given that a set of features has been sampled from a presented letter, it is the task of the decision process to use this information to select a response.

There are two central mechanisms that share, to a greater or lesser degree in the various models, the fulfilling of this task. The first is the acquisition of a measure of similarity of the sensory representation with each of the letters in memory. This acquisition is accomplished by matching features of the sampled set with those making up the memory letters. Some function of these comparisons determines a so-called confusion or candidate set of possible responses, and at this point, the second mechanism comes into play.

This second mechanism must select a "winner" from the members of the confusion set. When more than one alternative is contained in this set, it is assumed that the final choice occurs according to a set of guessing biases.

The multicomponent model, the hit-ratio model, and the feature-distance model are based on the idea of exhaustively testing the features belonging to the memory alternatives against the features appearing in the sampled feature set. They also allow for nondegenerate confusion sets (containing more than a single member). However, the sequential feature-testing model assumes a probabilistic, self-terminating feature-testing sequence and predicts that the process always ceases when exactly one letter alternative remains. That is, the latter model never allows a confusion set of any size other than one.

In the most general model one could imagine, each specific set of sampled features could, when matched against the features of the letters in memory, have a unique similarity measure with each memory alternative. However, to promote testability it seems reasonable to group the kinds of similarity that can arise into three main categories: (a) features overlapping in both the extracted set and a memory letter (constituting evidence for the memory letter), (b) features in the sampled set but not in the memory letter (ghost features, evidence against the memory letter), and (c) features not in the extracted set but in the memory letter (lost features, also evidence against the memory letter). If we call the sampled feature set F and a particular memory-response alternative R, then we can refer to features of type (a) as $F \wedge R$, to those of type (b) as $F \wedge \bar{R}$ and those of (c) as $\bar{F} \wedge R$. Similarly, $n(F \wedge \bar{R})$ will refer to the number of features contained in F but not R and so on.

Multicomponent Model

The multicomponent model by Rumelhart (Note 2) assumes that any alternative that lacks features that are present in the sampled feature set is automatically rejected; that is, if $n(F \wedge \bar{R}) > 0$, the memory alternative R is not admitted to the confusion set. On the other hand, a constant number c specifies the greatest number of features that a memory alternative might have that are lacking in the sampled feature set; that is, $n(\bar{F} \wedge R) \leq c$. Therefore, given any alternative R such that there exist no features in the extracted set F not also existing in R and that, at most, c features in R are missing from F, then R enters the confusion set of F, $C(F)$. As seen in the section on models of the feature-sampling process, the Rumelhart sensory phase posited that ghost features never occur; the decision process then is consonant with that assumption.

After $C(F)$ has been determined, the decision is made from the members of this set according to a set of response biases. The biases, that is, the guessing probabilities associated with a particular confusion set, were assumed by Rumelhart to be proportional to the relative magnitude of a member's Bayesian, or a posteriori, probability of having occurred as the stimulus. This latter assumption is inappropriate in the present experiment where we expect the biases to vary with payoff condition, but the a priori presentation probabilities are equal and constant. We assume in this and all other decision models that the probability of deciding on alternative

R_j given confusion set C is just the ratio of the response bias associated with R_j to the sum of all the biases of the alternatives in C. It can be assumed without loss of generality that the biases add to 1 across all the alternatives in the entire stimulus alphabet.

Rumelhart (Note 2) has fit the multicomponent model, the all-or-none model, the overlap model, and the choice model to a confusion matrix generated by six capital letters of the alphabet. The multicomponent model described the data appreciably better than did the all-or-none model and slightly worse than did either the overlap or the choice models. Rumelhart and Siple (1974) have extended the multicomponent model to make predictions about word recognition. The present decision part of the model has four parameters: the guessing bias parameters and the value of c.

Hit-Ratio Model

Several versions of a model, based on the assumption that recognition uses distinctive feature-lists devised from the universe of responses to process stimulus-evoked icons, were recently tested by Geyer and DeWald (1973).

The decision phase of the model, with which we are concerned here, compares the given feature set with the feature lists in memory that correspond to possible letter responses. As with the multicomponent sensory model, it is postulated by these investigators that ghost features do not occur in the absence of a noise background. However, memory alternatives with ghost features are not necessarily excluded from the confusion set as happens in the multicomponent model.

The measure of similarity between the extracted set of features and a memory letter is a hit ratio, h, where in our language

$$h = \frac{n(F \wedge R) - n(F \wedge \bar{R})}{n(R)}. \qquad (5)$$

Note that ghost features $(F \wedge \bar{R})$ have a strong negative effect, reflecting our belief in their lack of occurrence, and presumably, the observer's use of the fact in his or her decision making. Too, an indirect penalty is paid by memory alternatives possessing features not in the sampled set $(\bar{F} \wedge R)$ by way of the denominator of h, which is the total number of features in the response alternative R. One interesting difference between the multicomponent decision model and the hit-ratio model is that in the former, the overlapping features affect the decision only indirectly by being related to the number of features that can have been lost from R in that model, but in the latter model, the overlap comes in explicitly in the numerator of h.

The best-fitting version of the model as found by Geyer and DeWald (1973) contained an upper and lower threshold. From among memory alternatives with h values above the upper threshold, that memory letter having the maximum h was selected. If several alternatives obtained that maximum h, the observer used his or her guessing biases in deciding among them. If all computed h values fell below the lower threshold, it was predicted that the observer guessed from all possibilities using his or her response-bias probabilities, with probability of response R_j equal then to the bias B_j. When the greatest values of h lay between the upper and lower thresholds, it was posited that some fixed proportion of the time, the observer resorted to the below-threshold guessing strategy but that with one minus that proportion, he or she used the maximizing strategy. Thus, in the basic hit-ratio model, the confusion set depends on the threshold levels, on the one hand, being the set of alternatives reaching the maximum h value, and on the other, being the entire set of response possibilities.

In the present study, only one threshold was used for the sake of parsimony. However, two versions of this basic model were tested, both containing four parameters. The first, called simply the *hit-ratio* model, assumes that the decision process guesses with bias among the alternatives having hit ratios greater than the threshold and that it guesses among all alternatives when no hit ratio is greater than the threshold. The second version, the *maximum hit-ratio* model, picks that alternative associated with the maximum hit ratio if that ratio is above the threshold. If more than one alternative reaches the maximum hit ratio above the threshold, guessing occurs among those alternatives. Again, if none of the hit ratios are above the threshold, guessing takes place among all the possibilities. In the next section, we move to a model founded on the concept of distance.

Feature-Distance Model

In formulating this model, we took as a starting point the idea that the measure of dissimilarity between the extracted feature set and a memory letter might be a metric computed on the basis of the feature matches. It seemed natural, from this point of view, to formulate a pattern space made up of five dimensions, one dimension for each of the features that helps make up the letters used in the present study. Each memory letter and each extracted feature set has a value of 1 or 0 on each dimension depending on whether it contains that feature or not. Thus, any memory letter, as well as each extracted feature set, can be represented by a vector x extending from the origin to the point, $x = (x_1, x_2, x_3, x_4, x_5)$, where $x_i = 0$ or 1 characterizing the absence or presence of the ith feature. It is, of course, assumed that the dimensions are orthogonal to one another.

The distance between any extracted feature set and a memory letter is calculated as the sum of absolute differences on the various feature dimensions; this calculation is equivalent to computing the city-block distance between points (or vectors) in the simple pattern space outlined above. This measure of distance equals $n(F \wedge \bar{R}) + n(\bar{F} \wedge R)$.

The first variant of this model, which we refer to simply as the *feature-minimum-distance* model, responds deterministically in all cases with the letter response that is a minimum city-block distance from the extracted feature set. If however, there exist two or more letters in memory that yield the minimum distance on a trial, the system responds probabilistically, so that the frequency with which the observer chooses a particular one of the minimum-distance letters is proportionate to the relative magnitude of its response bias (in relation to the biases of the other minimizing letters). Hence, in this model, the observer only guesses if two or more memory letters are closest and equally close to the extracted feature set. This model, having three parameters, is generically similar to the "minimum distance classifier," discussed by Nilsson (1965) and others. However, the fact that "ties" on the minimum distance may occur and have to be decided by guessing, prevents its representation as a true minimum-distance classifier or indeed, as any linear discriminant function.

We call the second variant of the basic model the *feature-minimum-distance model with cutoff* because it adds a fourth parameter (in addition to the three free guessing bias parameters) that acts as a cutoff, d, so that any memory letter being more than d distance away from the extracted set point is deleted as a possible response. If none of the four distances are less than d, then the subject responds according to his or her biases. If there is at least one distance less than d, then this model operates exactly as the feature-minimum-distance model does.

The third form of the distance model, referred to as the *feature-distance cutoff bias* model, disregards memory letters with distances greater than d, but also supposes that the observer chooses among all letters with distances less than d according to their relative bias values, independent of which letter is associated with the minimum distance. It also possesses four parameters.

All variations of this model obviously give equal weight to lost and ghost features in terms of increasing the distance from the extracted set of features from a memory letter. In addition no direct weight is given the number of features overlapping the extracted set and a memory letter.

All of the models discussed so far share the postulate that all the features in the extracted set are tested for their presence or absence in all letters in the memory set. In this sense, the information in the extracted feature set is used exhaustively in the recognition process. Although models of this nature often suppose that the feature processing is done in parallel (as in the Pandemonium model; Selfridge, 1959), conceivably the feature set could be matched against all of the memory letters at once or one at a time, and similarly, the extracted features may be compared with the features of a single memory letter in parallel or serially.

Sequential Feature-Testing Model

In contrast to the above models, this model assumes self-termination; that is, testing features from the memory set against the extracted set does not have to exhaust all possible feature comparisons. The particular

Table 1
Model Characteristics of Feature Comparison

Reject R if $n(F \wedge \bar{R}) > 0$ or if $n(\bar{F} \wedge R) > c$	Weight $F \wedge \bar{R}$ and $\bar{F} \wedge R$ features equally	$F \wedge R, \bar{F} \wedge R, F \wedge \bar{R}$ features all important but with unequal weighting
Multicomponent model	All feature-distance models Sequential feature-testing model	All hit-ratio models

features selected for testing might be dependent on what features have previously been tested and whether they were found in the extracted feature set or not. A mechanism of this sort would have no purpose in the case of exhaustively processing features. The sequential feature-testing model has these properties. A set of weights is assumed to be attached to the set of features making up the memory letters, with one weight associated with each single feature. However, in the event that some features are in all the letters, they may be ignored; it is presumed in this model that the feature-testing mechanism is efficient in not wasting effort on nondiscriminative features.

The first feature, say feature i, to be tested against the extracted set is selected from the total set with probability,

$$\frac{w_i}{\sum_{j=1}^{n} w_j}, \tag{6}$$

where w_i is the weight associated with feature i and n is the total number of discriminative features. This selected feature is checked for its presence in the extracted set; if present, all memory letters containing it are retained, and all memory letters not possessing it are deleted from the set of possibilities. If this feature is absent from the extracted set, all memory alternatives possessing it are deleted from the set of possibilities, and those not containing it remain. The features remaining in the still viable memory alternatives form the pool for the next feature selection and feature k is chosen for testing with probability,

$$\frac{w_k}{\sum_{\substack{j=1 \\ j \neq i}}^{n} w_j}. \tag{7}$$

When all but one letter alternative have been excluded, the choice is completed. As noted, the process is self-terminating because not all of the extracted features need be matched against memory features. It is also contingent, with feature testing depending on the outcome of previous tests. As used in the present application, it has three parameters. As formulated here, the model is serial in nature and is in the spirit of such models as elementary perceiver and memorizer (EPAM; Feigenbaum, 1963).

The serial property is not necessary, as an exponential parallel model can give exactly the same predictions. The weights could be interpreted as exponential rate parameters, which then determine, along with the elimination mechanism, the features selected for testing. On the other hand, parallel models distinct from the serial model might result if the matching rates (weights) had different values depending on whether the tested feature was or was not in the extracted set (Townsend, 1976).

Tables 1 and 2 summarize the feature comparison and confusion set generation properties of the decision models.

In concluding our discussion of the feature-decision models, we mention briefly the possibility of hierarchical processing at different levels and the potential importance of relational or gestalt structure as opposed to feature combinations per se (see, e.g., Massaro & Schmuller, 1975). The present models could be modified to handle hierarchical processing by including in the set of features primitive features, such as size and gross shape, as well as the finer-level features, such as "a short horizontal line in the middle of the letter." A further modification would be to make, when necessary, the sampling probabilities of the finer-level features 0 except when preceded by sampling of the grosser

Table 2
Confusion Set as Determined by the Models

Set of alternatives with similarity to sampled feature set greater than some fixed number	Set of alternatives with maximum similarity to sampled feature set on a given trial	Set of alternatives with maximum similarity to sampled feature set, and this maximum is greater than some fixed number
Multicomponent model	Feature-minimum-distance model	Feature-minimum-distance with cutoff model
Hit-ratio model		Maximum hit-ratio model
Feature-distance cutoff bias model		

features. Thus, suppose overall shape extraction always precedes straight line feature extraction. Then the probability that the straight line feature i is sampled is greater than 0 only if shape has already been sampled. One may then posit confusion sets from which informed guesses are made just as in the models described above.

The spirit, then, of the present models is not antagonistic to hierarchical processing, but it may be helpful, as in this study, to substantially lessen the possibility of hierarchical processing by using a simplified synthetic font. In this way, something may be learned about simpler levels of feature analysis. In those instances where it is suspected that several levels of processing are performed, it would be best to demonstrate through actual quantitative fits that simpler models are inadequate. An analogous example has been the falsification in several cases of the all-or-none model as compared with models based on interstimulus similarity such as the choice and overlap models.

With regard to the question of gestalt relations, there seem to be two main aspects of especial importance here. One aspect relates to the fact that the simultaneous detection (sampling) of two (or more) features, rather than the separate detection of either one, may be critical. For example, to successfully discriminate lowercase b, p, q, and d from one another, it is necessary to extract both the vertical line as well as the curved feature (ignoring potential cues from seriphs, etc.). The structural relation poses no problems for the above models because whenever any particular combination of features is unique to a given letter, the correct letter will typically achieve a higher similarity measure and will be reported more often. For example, to consider an extreme instance, a sampled set that is unique to a single character will always lead to the correct response if ghost features are never sampled and if the decision maker consistently rejects any memory alternatives containing ghost features; all incorrect alternatives will be rejected for not having one or more of the features in the extracted set, whereas the correct alternative will be accepted. A model satisfying these criteria would be the multicomponent model with an upper limit of c that is large enough to permit inclusion of the correct memory alternative in the confusion set.

The other aspect of interest is that gestalt combinations of features should tend to be extracted together, that is, to have a high positive extraction correlation (some might demand a correlation coefficient of 1 for a truly gestalt feature-combination). This possibility can be tested in the present data on feature report.

Experimental Conditions

It may be recalled that the recognition experiment involved four synthetic stimulus letters (A, E, F, H) constructed from line segments of equal length; these line segments were operationally defined features. The purpose of the equal lengths was to rule out line length as the cause of any differences in sampling probabilities. As indicated in Figure 3, the line segments were physically connected, and no lines protruded from the natural form of the letter. In this regard, the present characters may be more "natural" than are synthetic letters that are made up of discontin-

uous lines and of lines that extend beyond the usual boundaries of letters (cf. Hubert, 1972; Schulze, Baurichter, Gerling, & Grobe, 1977; Rumelhart, Note 2).

The four stimulus letters were presented with equal probability, and the payoffs were varied on A and E versus F and H. The observers always gave both a feature report as well as a letter response on regular experimental trials, but a control condition was included in which only letter responses were elicited.

Method

Subjects

Three naive undergraduates with 20/20 vision served as paid observers. The pay was the prevailing minimum wage plus the bonus described later.

Apparatus

A Gerbrands two-field tachistoscope (model T-2B) was used to present the stimuli.

Stimuli were four letters, A, E, F, and H drawn in black ink with the aid of a template, and were presented on 5- × 8-in. white index cards, one letter per card. The four letters were constructed from five line segments of equal length and were presented according to the font illustrated in Figure 3. A prestimulus fixation field was described by a set of four dots that were arranged as the corners of a square with the letter in the center. The four dots were present at all times on the screen except during the brief intervals of stimulus presentation. The fixation field on any one side subtended an angle of about 2°, and a single letter, an angle of about ¼° at the observer's eye. The luminance of the stimulus was 7.1 fl. Responses were given verbally and were recorded by hand on recording sheets.

Procedure

Observers were instructed to give two responses (forced choice) to every stimulus presentation, namely, (a) a list of perceived line segments, or features (numbered 1–5 as in Figure 3), and (b) the name of the letter presented. This response procedure was followed in two within-session experimental conditions. In the FH condition, observers were paid 1¢ for each correct F or H response and ½¢ for each correct A or E response. In the AE condition, observers were paid 1¢ for each correct A or E response and ½¢ for each correct F or H response. Nothing was paid for incorrect responses. However, the observers were told the correct response on these trials as well. The trials were observer-paced following the experimenter's announcement of the applicable payoff condition for the upcoming trial, and the intertrial interval averaged about 10 sec. A warning click was sounded 500 msec before the stimulus presentation.

The observers were already much practiced in the task due to their participation in earlier experiments. Before the experiment began, the durations of the stimulus were individually calibrated so that their probability of being correct was low but greater than chance (.25). The average accuracy of observers in the experiment was $.3 \leq P(\text{correct}) \leq .4$. One more day of practice preceded 20 experimental sessions, each of which was divided into 4 blocks with 56 trials per block, each of which consisted of an equal number, 28, of FH and AE trials. In each group of 28 trials, an equal number of each stimulus was presented. Presentation of each stimulus letter with each condition was random within the foregoing constraints.

A control condition of six blocks wherein observers listed only letters was interspersed among the other trial blocks. Control blocks were randomly placed throughout the 20 experimental days with the constraint that no more than one a day be chosen.

Results and Discussion

The Control Condition and the Two-Phase Markov Assumption

The letter responses and the associated 4 × 4 confusion matrices would be of import even if they differed from results obtained in the absence of feature reports.[1] However, it may be of interest to ascertain whether the feature reports greatly perturbed the letter-report process.

Inspection of the control data (results of trials on which observer responded with only letter names) reveals that confusion matrices from the control condition appear very similar to corresponding matrices from the experimental condition. To provide a more rigorous test, the best of the descriptive models, the good-fitting choice model, was fit to the control matrices. As is seen below, in all instances, choice predictions were statistically indistinguishable from the control data ($p < .05$).

Choice-model parameter estimates averaged over all observers are listed in Table 3 for both experimental (feature reports given) and control (no feature reports given) conditions. Comparison of these estimates within payoff conditions shows that in only 2 out of 20 instances (M_{AH} and M_{FE} in the AE condition) were differences in the parameter estimates at all sizeable. It appears that the listing of features did not drastically affect the

[1] To conserve space, the quite cumbersome feature-report matrices are not given in this article. However, they may be obtained on request from the first author.

Table 3
Averaged Chi-Square Parameter Estimates of Descriptive Models

Condition	Choice						Overlap						All-or-none			
	Sensory parameters															
	M_{AH}	M_{AF}	M_{AE}	M_{HF}	M_{HE}	M_{FE}	E_{AH}	E_{AF}	E_{AE}	E_{HF}	E_{HE}	E_{FE}	P_A	P_E	P_F	P_H
Experimental																
FH	.70	.65	.58	.64	.49	.71	.41	.40	.19	.43	.21	.30	.06	.10	.13	.22
AE	.68	.67	.66	.44	.44	.71	.32	.29	.39	.21	.36	.40	.07	.16	.11	.19
Control																
FH	.73	.69	.51	.68	.57	.75										
AE	.79	.65	.67	.50	.45	.50										
	Bias parameters															
	b_A	b_E	b_F	b_H			g_A	g_E	g_F	g_H			h_A	h_E	h_F	h_H
Experimental																
FH	.17	.15	.35	.33			.18	.15	.34	.33			.19	.14	.37	.30
AE	.32	.29	.19	.20			.31	.29	.19	.21			.37	.28	.18	.16
Control																
FH	.16	.16	.34	.34												
AE	.28	.30	.21	.21												

Note. FH and AE = experimental conditions in which stimulus letters F, H and A, E received the higher payoffs for correct responses, respectively.

letter reports. Without doubt, experimental circumstances, and in particular, exact stimulus and response conditions are often critical in determining the type of underlying psychological process that is taking place (see, e.g., Garner & Haun, 1978; Garner & Morton, 1969). Yet, thankfully there appear to be cases where some useful invariance of processing mechanisms occurs.

Next, consider the conditional relation between the feature reports and the letter reports. The feature reports may yield valuable information about what the observer is seeing when a letter is presented, even if the letter reports, when conditionalized on the reported feature set, are not independent of the stimulus letter presented. However, the combined data on feature report- and on-letter report allows, in principle, a provisional test of the traditional Markov assumption relating the hypothetical feature sample to the following letter decision. If the "true" set of features (those actually used by observers to identify alphanumeric characters) is known, then the assumption of separate sensory and decision processes implies a statistical independence of the data on feature and letter report. That is, given any specific reported feature set, the response to it on any trial should be independent of the stimulus presented, and therefore, the entries in the confusion matrix associated with that feature set should be completely predictable given the marginal frequencies. Because this test requires a priori knowledge of the "true" set of features, a statistical dependency could be due to a use of an incorrect feature set, the failure of our assumption of separate feature-sampling and decision phases, or the failure of the Markov assumption relating the feature sample to the letter decision. In any event, independence would support the tenability of the assumption and would also indirectly tend to increase the viability of the presently used feature set.

The matrices that are formed by conditionalizing on the reported feature set are of a large number because there are 32 possible feature sets, and the matrices that can be formed by conditionalizing on the stimulus letter or the response letter have many cells (32×4). There is, therefore, some difficulty encountered in interpretation of chi-square values due to a number of cases where more than 20% of the cells have associated expected values of less than 5. Also, there appears to be no theoretically meaningful way to combine data to alleviate this problem. These particular analyses must therefore be considered very tentative. However, we believe the technique is of sufficient interest and the findings, at least suggestive enough to justify a brief look at the results.

First, consider a chi-square test of independence of the feature sets and responses when we conditionalize on the stimulus letter presented. This figure should provide a baseline guide because independence would argue that no relation exists between the reported features and the reported letters. The smallest value obtained for any of the three observers was 1,738 ($df = 264$) in condition AE, and the largest was 4,916 ($df = 219$) in condition FH. The disparity in degrees of freedom was due to differing numbers of missing cells. The normal standard deviates corresponding to these are $z = 36$ and $z = 78$, respectively. The hypothesis of no relation between feature and letter reports is patently false.

Next, does the stimulus letter lead to a feature set as hypothesized, and thence to a response, or does the letter decision follow the stimulus immediately, and the observer then picks a feature set (to comply with the experimenter's instructions) based on that response? Although other alternatives are possible, these two seem most interesting and, in principle, open to test. Our results here are, unfortunately, inconclusive. Testing the traditional assumption by conditionalizing on the reported feature set supported the null hypothesis of statistical independence at the .05 level for five of the six observer experimental-condition cases. Four of the chi-square values were essentially equal to the degrees of freedom. This finding is in line with the theory (i.e., the two-phase Markov assumption).

On the other hand, a similar result was obtained when we conditionalized on the letter response, although here only four of the six chi-square values were statistically nonsignificant, and the normal approximating standard deviates tended to be somewhat larger.

Table 4
Empirical Confusion Matrices for Two Experimental Conditions Averaged Over all Observers

Letter stimulus	A	E	F	H
	FH condition			
A	.245	.122	.306	.326
E	.148	.227	.385	.239
F	.160	.130	.460	.250
H	.164	.096	.280	.460
	AE condition			
A	.417	.237	.157	.190
E	.312	.400	.171	.118
F	.303	.305	.272	.120
H	.329	.213	.137	.321

Note. FH and AE conditions = the experimental conditions in which the stimulus letter F, H and A, E received the higher payoffs for correct responses, respectively.

The major reason that we cannot definitively discriminate between these two latter alternatives in the present case appears to be the relatively low accuracy in the present experiment (which was necessary for certain other tests). That is, as perceptual accuracy falls, the relation between the stimulus and the features is decremented, or alternatively, the relation between the stimulus and the letter response diminishes, and the two theoretical possibilities tend to make similar predictions and therefore become difficult to distinguish. We hope that future experiments with more varied performance levels will permit investigators to use the suggested technique.

Overall Descriptive Fits

Table 4 shows empirical confusion matrices for the two experimental conditions averaged across all observers, whereas Table 5 contains the confusion matrices for the three individual observers.

In the interest of space, Table 6 shows only the averaged theoretical predictions of the descriptive models. However, as expected, the choice model provided the best fits to the data for every observer. As is evidenced in the averaged tables, the empirical matrices were basically compatible with those predicted by choice theory (e.g., the average chi-

Table 5
Empirical Confusion Matrices for the Three Individual Observers in the Two Conditions

	FH condition				AE condition			
Stimulus letter	A	E	F	H	A	E	F	H
	Observer A							
A	.212	.098	.383	.307	.371	.364	.171	.093
E	.128	.143	.413	.316	.310	.457	.143	.090
F	.093	.072	.468	.368	.283	.447	.209	.062
H	.120	.074	.376	.431	.347	.311	.157	.185
	Observer B							
A	.252	.152	.264	.331	.395	.207	.179	.219
E	.148	.269	.364	.219	.253	.422	.193	.131
F	.193	.188	.440	.179	.288	.302	.269	.140
H	.204	.140	.220	.436	.313	.212	.114	.360
	Observer C							
A	.271	.117	.271	.340	.483	.138	.121	.257
E	.169	.269	.379	.183	.374	.319	.176	.131
F	.193	.131	.471	.205	.338	.167	.338	.157
H	.169	.074	.245	.513	.327	.115	.141	.418

Note. FH and AE conditions = the experimental conditions in which the stimulus letters F, H and A, E received the higher payoffs for correct responses, respectively.

square value over the six confusion matrices was $\chi^2[3] = 5.20$, $.10 < p < .20$). In only one instance was an empirical matrix compatible with the all-or-none theoretical prediction ($\chi^2[5] = 7.48$, $.2 < p < .1$; for the many statistically significant instances, the average was $\chi^2[5] = 28.458$, $p < .001$) and in no instances were overlap predictions statistically the same as the empirical matrices (the average was $\chi^2[3] = 134.037$, $p < .001$). For comparison with earlier papers (e.g., Townsend, 1971a, 1971b), the averaged summed squared deviation for the choice model was .0036, for the all-or-none model, .0133, and for the overlap model, .1001.

As predicted, the overlap model was strongly falsified relative to the choice model. Thus, the nonparametric overlap prediction mentioned earlier, that

$$\sum_{\substack{i=1 \\ i \neq j}}^{N} c_{ij} \leq c_{jj} \tag{8}$$

for $1 \leq j \leq N$, can be seen to be violated in every single individual confusion matrix. Computer procedures to force the necessary row sums to equal 1 often allowed the overlap predictions to violate this prediction (see, e.g., column F under FH condition within Table 6) but without significantly improving the fits. That this falsification is due to the higher interletter similarity relative to earlier studies is supported by comparison of similarity indices computed on both our and Rumelhart's (Note 2) stimuli. The average value of M, the choice model's similarity parameter, was .71 here and .15 for Rumelhart. Further, both the physical overlap mea-

Table 6
Averaged Theoretical Predictions of Descriptive Models Plus the Best Fitting Feature-Sampling and Decision Models (Multicomponent Model and Eight-Parameter Feature-Sampling Model)

	FH condition				AE condition			
Stimulus letter	A	E	F	H	A	E	F	H
				Choice				
A	.245	.117	.316	.322	.416	.252	.156	.175
E	.154	.227	.379	.240	.297	.399	.180	.124
F	.150	.137	.460	.253	.303	.296	.272	.128
H	.168	.095	.277	.460	.344	.206	.129	.321
				Overlap				
A	.387	.085	.262	.266	.569	.193	.111	.127
E	.121	.352	.323	.205	.236	.540	.131	.092
F	.107	.093	.599	.200	.250	.241	.410	.100
H	.122	.067	.226	.586	.290	.171	.111	.127
				All-or-none				
A	.244	.126	.348	.283	.414	.268	.167	.151
E	.167	.225	.335	.273	.312	.348	.152	.137
F	.161	.117	.459	.263	.329	.256	.270	.145
H	.145	.105	.291	.459	.300	.234	.146	.320
				Multicomponent plus six-parameter feature sampling				
A	.244	.106	.320	.330	.425	.221	.163	.191
E	.147	.226	.394	.232	.289	.398	.202	.112
F	.163	.148	.435	.255	.327	.317	.229	.127
H	.168	.096	.276	.459	.320	.216	.148	.316

Note. FH and AE conditions = the experimental conditions in which the stimulus letters F, H and A, E received the higher payoffs for correct responses, respectively.

Table 7
Summed Squared Deviations of Five Feature-Sampling Models Predicting a 4 × 32 Matrix

Average across observers	Model 1 ($\alpha, \gamma = 0$)	2 (α, γ)	3 ($\alpha_1, \ldots, \alpha_5, \gamma = 0$)	4 ($\alpha_1, \ldots, \alpha_5, \gamma$)	5 ($\alpha_1, \ldots, \alpha_5$ $\gamma_2, \gamma_3, \gamma_5$)	Base value Uniform 1/32
FH condition	.342	.2783	.2652	.1674	.1266	.4697
AE condition	.3465	.2113	.3136	.1391	.0945	.3730

Note. FH and AE conditions = experimental conditions in which stimulus letters F, H and A, E received the higher payoffs for correct responses, respectively; α = probability of sampling a feature that exists in the stimulus letter; γ = probability of sampling a feature that does not exist in the stimulus letter.

sure of letter similarity (Rumelhart: .64 vs. Townsend: .47; see Townsend, 1971b) and the measure $n(S_A \wedge S_B) - [n(\bar{S}_A \wedge S_B) + n(S_A \wedge \bar{S}_B)]$ for any two stimulus letters S_A and S_B (Rumelhart: 1.0 vs. Townsend: −1.27) substantiated the higher similarity in the present study. Not surprisingly, the all-or-none model was also not able to handle the situation of high psychological similarity.

Table 3 contains the parameter estimates of the three models averaged across FH and AE conditions and across observers. Note that in all cases the bias parameters make the appropriate changes across payoff conditions. Specifically, for all three models, bias parameters associated with responses A and E increase from the FH to the AE. Likewise, biases for H and F decrease from the FH to the AE.

Because the all-or-none model and the overlap model clearly cannot explain the data, we do not expect their stimulus-related (or sensory-related) parameters to remain constant across payoff conditions, but we might hope for this constancy from the good-fitting choice model. Although four of the similarity parameters are relatively invariant, M_{AE} and M_{HF} show rather more variability than we should like. It is interesting to note from Table 3 that the effect of all the stimulus-related parameter changes across bias conditions in the various models is to enhance the reporting of the more highly rewarded responses. Hence, the hypothesis that some alteration in the stimulus-related phase mechanisms actually occurred across bias conditions should at least be considered. If these parameters do actually represent some perceptual process, then we might expect to find corresponding parameter changes across payoff conditions in the feature-processing models. Such findings would place added confidence in any psychological interpretation that these descriptive models might yield. We shall return to this finding in the results section on feature sampling.

Substantive Models: The Feature-Sampling Process

In turning to separate models of the feature-sampling and decision processes, it should be kept in mind that in the past, specific recognition models have typically not performed quite as well as the best of the descriptive models. We hope to partially diagnose the reasons. To seek an answer, we should first examine the success of the feature-sampling models. Sensory models predicted a 4 × 32 confusion matrix. The four rows represent the four letter stimuli, and the 32 columns represent all possible combinations of feature lists. These matrices cannot test models by chi-square very efficiently because many of the 128 cells are empty; hence, the theoretical-observed squared deviations summed across a matrix were used instead. Table 7 shows summed squared deviations of the five sensory models studied with the empirical matrices. Also included is the summed squared deviation of a model that makes uniform predictions across rows (in this case, $\frac{1}{32}$ in every cell). This technique produces a base value of nonprediction above which any model possessing even the slightest predictive validity was expected to rise; it has been used in several previous studies

Table 8
Least Square Parameter Estimates of Most General Feature-Sampling Model

Average across observers	Parameters							
	α_1	α_2	α_3	α_4	α_5	γ_2	γ_3	γ_5
FH condition	.851	.522	.635	.727	.197	.284	.418	.087
AE condition	.858	.556	.686	.683	.307	.349	.488	.187

Note. FH and AE conditions = experimental conditions in which stimulus letters F, H and A, E received the higher payoffs for correct responses, respectively. α = probability of sampling a feature that exists in the stimulus letter; γ = probability of sampling a feature that does not exist in the stimulus letter.

(e.g., Geyer & DeWald, 1973; Townsend, 1971a, 1971b). Note that falsification of critical assumptions is more powerful when the model contains many parameters than when it contains only a few, contrary to what one occasionally reads in the literature.

The fits of the models were surprisingly poor. The simplest single-parameter model of Rumelhart (Note 2) and Geyer and DeWald (1973) did only slightly better than did uniform prediction, and even the eight-parameter model left a large summed squared error. A diagnosis for this result is taken up in the next subsection. In the remainder of the present discussion, some other overall characteristics of the feature reports are described.

Table 8 lists the average values across observers and within conditions of the eight parameters in the most general sensory model. Recall that alpha corresponds to the probability of reporting a feature that exists in the stimulus, whereas gamma represents the probability of reporting a feature that does not exist in the stimulus. Note that the gamma values are substantially greater than 0. This fact, together with the fact that the summed squared errors of models allowing gamma to deviate from 0 were smaller than those of corresponding models that did not, indicate that, in fact, observers did perceive ghost features, or features that did not exist in the stimulus. The conclusions were corroborated by calculating the marginal frequencies of ghost feature reports from the 4×32 matrices. This result is contrary to previous mathematically specified sensory-feature models. It may be that previous overall descriptive fits of these models to ordinary confusion matrices were adversely affected by the assumption that gamma equals 0. Perhaps the high-similarity, low-accuracy stimulus situation encountered by our observers was more conducive to noise infiltration than were some previous studies. In two studies further discussed below, Hu (Note 5) and Townsend, Hu, and Ashby (1980) again found evidence in favor of ghost features.

Garner and Haun (1978) found that brief tachistoscopic exposures are likely to lead to lost features, whereas information may be gained or lost when perceptual degradation occurs due to experimentally intruded noise. Further, as Garner (1978) noted, the Townsend (1971a, 1971b) data suggest a preponderance of lost feature confusions. Nevertheless, the present work indicates that featural information may be gained even in the absence of experimentally imposed noise. Certainly this outcome is compatible both with classical notions of signal detectability theory as well as with the well-known nontrivial spontaneous firing rates of sensory neurons.

Next, note that α_1 and α_4 (sampling probabilities for the always-present features) although larger than the alphas for the other features, do not approach 1, providing indirect evidence that the observers were conscientiously attempting to follow instructions in reporting their feature sample. Whether the greater sampling probabilities for these always-present features is due to sensory or bias mechanisms cannot be determined from the present data.

Table 8 also shows that α_3, α_5, and all values of gamma increase when shifting from the FH to the AE conditions, which implies that observers consistently reported features, whether or not they existed in the stimulus, more frequently when rewarded for correct

A or E responses than when rewarded for correct H or F responses. Note that F and H each contain three features; it is plausible that the observers may have lowered their "feature criterion" in the AE condition, favoring the stimuli with more features (A, E).

The alphas correspond to the probability of a hit, whereas the gammas correspond to the probability of a false alarm. This structure suggests the use of a signal detectability analysis to help explain the changes in alpha and gamma across the bias conditions. The results revealed no consistent change in d' (d' = difference between means of noise and signal and noise distribution relative to the standard deviation), but the decision criterion was lowered for all features from condition FH to condition AE. This finding agrees with the earlier noted change in the choice-similarity parameters across the bias conditions and is compatible with the hypothesis that they arise from decision-type influences on feature sampling and/or feature reporting.

Feature-sampling independence and invariance assumptions. It will be recalled that there are two main types of invariance plus the notion of sampling independence to be considered. First consider across-letter invariance for a given feature. Table 9 gives contingency tables of marginal response frequencies with which each feature was reported to each stimulus. Observers B and C are separated from Observer A because they revealed similar performance strategies that differed from that of Observer A.

The table entries are feature-report frequencies based on trials where the feature was contained in the letter, and hence, a frequency corresponds to alpha in the feature-sampling models. It may be seen that there is reasonable constancy for any given feature across the four letters, and a chi-square independence test of this assumption produced no significant differences at the .01 level for any of the 20 possible cases (10 for Observer A and 10 for Observers B and C). Further, the d' analyses mentioned earlier substantiated this result. Thus, we may conclude that for a given feature, neither sensory sensitivity nor bias varies greatly across the four letters used in the present study. This assumption corresponds to the assumption that α_i is constant across the four letters for $i = 2$, 3, or 5 (see Figure 3).

On the other hand, reading the rows of Table 9 shows that the individual features are reported with different frequencies, which is consonant with the estimates of the alphas in the preceding section. However, the largest range in d' values among the three information-bearing features and the two payoff conditions (Features 2, 3, and 5) was only .423 to .563, suggesting that the variation in

Table 9
Contingency Tables of Frequencies With Which Each Feature Was Reported to Each Stimulus Averaged Over Three Observers

	FH condition					AE condition				
Stimulus letter	1	2	3	4	5	1	2	3	4	5
					Observer A					
A	363	210	283	341	—	337	190	344	256	—
E	369	—	274	324	60	390	—	331	262	174
F	369	—	253	321	—	394	—	341	247	—
H	362	211	—	335	—	373	213	—	280	—
					Observers B and C					
A	337.5	203	214.5	298.5	—	330	214	231	297.5	—
E	332.5	—	252.5	265.5	86	331.5	—	271	287	97.5
F	334	—	262.5	284.5	—	341	—	269	288.5	—
H	340.5	227	—	299	—	325.5	230	—	295.5	—

Note. FH and AE conditions = experimental conditions in which the stimulus letters F, H and A, E received the higher payoffs for correct responses, respectively; 1–5 = the five features making up the stimulus letters.

frequency of feature reports of Table 9 was due to distinct biases on the three features rather than differential sensory sensitivity. It is interesting to consider the higher report frequency of Feature 3 (at the top) than of Features 2 and 5 (on the right and bottom) and the more frequent reporting of Feature 1 (on the left) than of 4 (in the middle). This report is compatible with a number of other studies of native English readers. At least in our results, we have the suggestion that this difference is due to criterion factors rather than strictly to sensitivity factors. Note also that the magnitude of α_i generally is ordered according to the frequency with which feature i appears in the four letters.

Feature-sampling independence is an important assumption because it allows the probability that any particular subset of features is sampled to be written as a multinomial combination of the separate feature probabilities. Because across-letter invariance was supported above and across-feature-sampling variability and ghost features were allowed in the more general of our feature sampling models, sampling independence is immediately suspect as the potential culprit in the poor fits of those models to the feature reports. Examination of the data found widespread violations of independence. For instance, if the feature pair i–k is reported more frequently than is the pair j–k then the pair i–m should be reported more frequently than is the pair j–m (based on a specific stimulus letter); this prediction was frequently falsified.

The apparent failure of the assumption of sampling independence may be one reason that feature models have not performed better in the past and may prove an impediment to feature modeling if some solution is not found to the problem. Close analyses of the present data suggested an important cause of sampling dependence: Observers almost never reported feature combinations that could not have occurred by feature loss from one of the four stimulus letters. For instance, any feature combination including both Features 2 and 5 (e.g., ⅃) was virtually never "seen." That is, the observers behaved as though ghost features could not occur if it meant reporting such an "impossible" feature subset. It is seen later that this strategy is compatible with the letter-decision behavior of our observers. The exclusion of such feature combinations would be expected to severely damage sampling independence.

It is unclear at present whether the rejection by observers of "impossible" feature combinations is due to a fairly early filtering or gating out of these possibilities or, more likely, to a higher-order cognitive decision operation. Even if it is the latter, it may well be that such decisional effects occur in typical recognition studies prior to letter selection and thus also tend to cause violations of sampling independence there.

However, this factor is not sufficient to explain all the dependence found in the data. If feature-sampling independence held except for this effect, one should find independence among subsets not including the "impossible" combination. In fact, substantial violations are discovered even among such subsets. As an example, with stimulus letter A, the feature subset 2, 4 was reported 106 times, and 2, 3 was reported only 21 times, suggesting that $\alpha_4 > \alpha_3$, if sampling independence is true. However, 1, 3 was reported 159 times, and 1, 4, 105 times, implying contradictorily that $\alpha_3 > \alpha_4$.

Hubert (1972) reported evidence for sampling dependencies at all subset levels with alphanumeric types of characters. However, no decisional phase was included in the models, and group averaged data was tested. On the other hand, Schulze et al. (1977) discovered sampling independence, across-letter invariance, and across-feature variability when their synthetic features were widely separated. Their latter two results agree with our findings, but the first does not. With connected features, the sampling independence failed in the Schulze et al. data, and it appears from their tables that across-feature and across-letter invariance both fail in that condition, too (although the latter two assumptions were not directly tested). Thus, their more letterlike stimuli also produced violations of sampling independence. However, the types of features used in that study were quite different from those here so that cross-study generalization is difficult.

Wandmacher (1976) recently carried out a feature study with stimuli made up of either a single straight line or two lines connected

at an acute angle. His models contained a feature sampling as well as decision stage, but ghost features were not allowed. If two stimuli had the same number of features, then across-letter (actually across-stimulus, because letters were not used as stimuli) invariance was found, but feature-sampling probabilities went down in symbols with larger numbers of features. Because our letters had about the same number of features equal in length, his latter results are not incompatible with the present findings or with those of Taylor (1976). He also, like most other investigators (e.g., Geyer & DeWald, 1973; Schulze et al., 1977; Rumelhart, Note 2) and ourselves, found it necessary to posit unequal sampling probabilities (i.e., $\alpha_i \neq \alpha_j$, for $i \neq j$) for the separate features. The hypothesis of sampling independence could not be rejected on the basis of his experiment and within the context of his models.

Townsend et al. (1980, 1981) studied the recognition of two orthogonally arranged straight line segments, operationally defined as features. Both, either, or neither of the features were presented. Because all combinations of the features were possible, including the blank stimuli, it was feasible to test for a strong version of sampling independence in which it was assumed that a feature is reported if and only if it is perceived, with no additional intervening decision process. Sampling independence was obtained where decisional influences were reduced to a minimum.

In contrast, when an alphabet of 16 symbols, constructed from all combinations of four straight and curved lines, was presented tachistoscopically to four observers, severe dependencies were discovered (Hu, Note 5). Because no feature combinations were "impossible," that explanation is unable to account for the lack of sampling independence.

Thus, mounting evidence suggests that a reasonable approximation to sampling independence of features, defined as line segments, can be found in very simple circumstances but that it readily disappears with even marginally complex stimuli. A plausible next step is to assume that some combination of lines form gestalt features, which are themselves independent. For instance, perhaps the vertical line and the top horizontal line form a right angle "feature" that is independently sampled with the other horizontal lines. All such possibilities were tested in the present data; none succeeded. It was not feasible to test all such combinations of "feature" possibilities embedded in a full model containing a decision phase in the above-mentioned Hu (Note 5) study, but they were tested in a direct, unbiased report model (i.e., it is assumed an observer reports exactly what he or she samples) with very poor fits. More abstract topological "features," such as roundness, jaggedness, vertical linearity, and so on suggested in other contexts (e.g., Kuennapas, 1966; Townsend, 1971a, 1971b), are not sensible in the present conditions.

Of course, if letter recognition does not occur at all by way of feature sampling, failure of sampling independence is not surprising. The preceding and other experimental studies provide evidence that at least certain assumptions of feature sampling are satisfied in some experimental contexts, but certain other possibilities are brought up in the final section of the article.

Substantive Models: The Decision Process

Models of the decision process predicted a 32×4 confusion matrix where the 32 stimulus rows represented all possible combinations of features, and the four response columns, the letter responses A, E, F and H. Table 10 shows the summed squared deviations of the theoretical predictions from the empirical confusion matrices. It may be remembered here that all models had four parameters except the sequential feature tester and the feature-minimum-distance model, both of which had three.

It can be seen that the models performed much better than the uniform prediction index, and in several cases, the fits were almost indistinguishable from the data. All models also tended to exhibit appropriate shifts of their bias parameters across payoff conditions. The maximum hit-ratio model of Geyer and DeWald (1973) provided the best average fit to Observer A whereas the multicomponent model of Rumelhart (Note 2) performed best for Observers B and C. Table 11 shows the parameter estimates of the models for the various observers and conditions. An examination of these two tables

Table 10
Summed Squared Deviations of Decision Models Fit to 32 × 4 Decision Matrix

Observers and conditions	Model							
	Feature-minimum-distance	Feature-distance bias	Feature-minimum-distance with cutoff	Hit-ratio	Maximum hit-ratio	Multi-component	Sequential feature-testing	Uniform (¼ in every cell)
Observer A								
FH	.333	12.545	.822	10.687	.791	2.01	.333	16.008
AE	10.54	7.801	7.328	14.29	4.98	9.456	10.54	13.794
Observer B								
FH	1.927	9.814	1.718	9.814	5.814	.666	1.927	12.485
AE	5.955	8.653	11.369	6.35	11.369	1.139	5.955	11.362
Observer C								
FH	1.166	14.445	1.10	11.79	1.10	.921	1.166	14.196
AE	3.827	9.189	2.55	10.756	7.428	2.36	3.827	12.299

Note. FH and AE = experimental conditions in which stimulus letters F, H and A, E received the higher payoffs, respectively.

is informative with respect to explaining the letter-decision behavior of the observers.

First, the generally good performance of the multicomponent model suggests that the observers tended to reject memory alternatives that possess ghost features relative to the sampled feature set. Scrutiny of the 32 × 4 feature-set versus letter-response matrices proved the following to be true: The observers essentially never responded with such an alternative. Note that although the observers received feedback that must have led to knowledge that ghost features could appear in their feature samples, their strategy of rejecting such alternatives still makes sense if feature losses outnumber ghost-feature occurrences. This condition held and was supplemented by another fact mentioned earlier, namely, that feature combinations that could arise only from ghost-feature sampling (e.g., ∃) never emerged at the feature-report level.

Second, consider the hit-ratio and feature-distance cutoff bias models (see Table 1). These models emphasize heavy guessing and, in addition, allow ghost-feature alternatives. They tended to perform quite poorly because of this ghost feature property but did improve somewhat in condition AE relative to condition FH. The other models, in contrast, tended to do a little worse in condition AE than in condition FH, although still better than the hit-ratio and feature-distance cutoff bias models in most cases. This result is explained by the increased guessing required in condition AE, the more favored letters A and E possessing more features. For instance, when Features 1, 3, 4 (an F) were sampled, the observers consistently responded F in condition FH, but in condition AE, they sometimes responded A or E as well.

Third, consider the parameter estimates of the multicomponent model, particularly the consistently large estimate of c and the sizable discrepancy between the A–E and the F–H biases. It will be recalled that c specifies the maximum number of features that may be lost from the stimulus during feature sampling. No stimulus in the present study contains more than four features, and thus, an estimated c of 4 means that this model is ignoring all lost-feature information. All response alternatives not containing any ghost features enter the confusion set regardless of the number of lost features they contain. This

Table 11
Least Square Parameter Estimates of Decision Models

Observers and conditions	Model						
	Feature-minimum-distance	Feature-distance bias	Feature-minimum-distance with cutoff	Hit-ratio	Maximum hit-ratio	Multi-component	Sequential feature-testing
Observer A	$B_A = 0$	Cutoff = 2.0	Cutoff = 3.0	Cutoff = .667	Cutoff = .0	Cutoff = 2.0	$W(2) = .093$
	$B_H = 0$	$B_A = .003$	$B_A = .009$	$B_A = .0008$	$B_A = .008$	$B_A = .00$	$W(3) = .364$
FH	$B_F = 0$	$B_H = .558$	$B_H = .787$	$B_H = .527$	$B_H = .788$	$B_H = .796$	$W(5) = 1.34 \times 10^7$
	$B_E = .999$	$B_F = .433$	$B_F = .195$	$B_F = .472$	$B_F = .195$	$B_F = .204$	
		$B_E = .006$	$B_E = .009$	$B_E = .0001$	$B_E = .009$	$B_E = .00$	
	.330	1.0	5.0	.667	.667	4.0	
	.174	.455	.298	.007	.205	.011	
AE	.165	.009	.027	.452	.129	.460	$W(2) = .175$
	.330	.009	.00	.509	.130	.498	$W(3) = .185$
		.528	.675	.032	.536	.030	$W(5) = 6.7 \times 10^6$
Observer B	.161	2.0	5.0	.50	.50	4.0	.419
	.357	.099	.186	.099	.233	.102	.463
FH	.322	.415	.359	.415	.308	.394	1.68×10^6
	.161	.386	.309	.386	.234	.403	
		.101	.146	.101	.225	.101	
		3.0	0	.667	1.0	4.0	.381
	.317	.236	.295	.214	.295	.196	.454
	.204	.205	.167	.290	.167	.304	3.39×10^5
AE	.162	.278	.290	.290	.291	.290	
	.317	.281	.247	.206	.247	.212	
Observer C	.237	1.0	5.0	.667	0	4.0	1.60
	.123	.065	.029	.011	.029	.014	.49
FH	.403	.123	.192	.192	.192	.227	1.67×10^6
	.237	.599	.724	.770	.724	.736	
		.213	.054	.026	.054	.024	
	.359	1.0	5.0	.667	.667	4.0	.610
	.123	.242	.288	.055	.284	.065	.423
AE	.178	.105	.211	.337	.169	.399	8.39×10^5
	.350	.278	.392	.583	.448	.515	
		.375	.110	.026	.10	.021	

Note. FH and AE = experimental conditions in which stimulus letters F, H and A, E received the higher payoffs for correct responses, respectively.

strategy naturally results in larger confusion sets and hence a greater reliance on the response-bias parameters. The relatively low accuracy characterizing this study tends to support this interpretation.

The large values of c are also responsible for the disproportionately larger F- and H-bias estimates that occurred in both payoff conditions. Note that the letters F and H do not contain any features not also found in the letter A. Thus with lost-feature information ignored, whenever an F or H are admitted to the confusion set, so also will be an A. But even in the AE condition, the great majority of the time observers responded to feature set 1, 3, 4 (an F) with F and to feature set 1, 2, 4 (an H) with response H. The multicomponent model mimics this behavior in the only way it can, by inflating the F and H biases.

From the above and similar analyses, we may surmise that: (a) the observers summarily rejected ghost feature alternatives; (b) they guessed among alternatives in their confusion sets with biases that changed appropriately across motivational conditions; and (c) if the sampled feature set was identical to F or to H, there was a preference given to responding F or H, respectively, despite the fact that A could have led to F or H (and E could have led to F) by feature losses alone.

With regard to the third point, it is not enough simply to institute an ability to respond with the closest (lost feature) alternative in the multicomponent model. Such a model was tried and fit no better than did the present multicomponent model, although the bias estimates were more reasonable in condition FH. Rather, future models for the present type of design will apparently have to include a set of special "confusion" states for the feature sets corresponding to perfect matches between the sampled feature set and a memory alternative. These special confusion states would then be associated with a special set of guessing-bias parameters favoring the perfect-match alternative.

Finally, we should mention that although none of the present models are equivalent in general, there are regions of the 32×4 space where two or more models can make equivalent predictions. Thus, the feature-minimum-distance and the sequential feature-testing models were able to mimic each other in the case of Observer A under condition FH. The reasons for this model mimicry can be found by considering the details of the observers' behavior relative to the model structure, but these considerations lie beyond the present scope.

Fit of A Combined Feature-Sampling and Decision Model to the Overall 4×4 Confusion Matrices

It is of interest to attempt to test the combination of the best feature-sampling model with the best present decision models.

Because including two rather than one parameter for ghost-feature occurrence did not lead to a large improvement in fits to the data on feature report, the six-parameter feature-sampling model (Model 4) was used to hold down the number of parameters somewhat (see Tables 7 & 8). This sampling model was joined with the multicomponent model and used to predict the individual 4×4 confusion matrices. The complete model possessed 10 parameters, one more than did the choice model.

First, the parameter estimates found with the separate data on feature and letter report were used. The predictions were not very good and, in fact, were substantially worse than were the choice-model predictions. Next, all the parameters were reestimated directly with the 4×4 matrices, and the average result is shown in Table 6. Although the fit is obviously not too bad (cf. Table 4), it still did not perform quite as well as did the choice model for any of the three observers. Given the generally satisfactory results of the multicomponent model to the letter reports conditionalized on feature set, it seems not unlikely that the lack of an excellent overall fit to the 4×4 matrices is due to violation of the sampling-independence assumption.

Summary and Conclusions

The major results and conclusions of the present study may be stated as follows:

1. The overt reporting of information that the observer believes she or he has seen has

been fruitful in testing, in conjunction with the model tests, hypotheses concerning feature processing and decision structure, and it appears to merit further investigation.

2. Our chi-square dependency analyses supported a relation between the observers' feature reports and their subsequent letter reports. The tests between the traditional Markov hypothesis (that the stimulus letter lead to an internal feature set, and then the letter decision is based on the feature set) and the alternative possibility (that, in this experiment, the observers immediately selected a letter response upon presentation of the stimulus and then reported a feature set based on this letter) were inconclusive. However, the confirmation of several notable findings of the present experiment by other recent studies (e.g., sampling dependence; see the fourth conclusion below) indirectly suggests that the natural sequential (Markov) assumption was appropriate here.

3. Among descriptive models that predicted the 4 (stimulus letter) × 4 (response letter) confusion matrices, the choice model again was superscendent, although some variation of the similarity parameters occurred across payoff conditions. The all-or-none model was expectedly falsified and the overlap model, as predicted, performed much worse than did the choice model because of the high interletter similarity among the four stimulus letters.

4. The assumption that the separate features are sampled in a probabilistically independent manner was invalid in the present and in some other recent studies. Further, no subsets of features were found that acted as gestalts but were independent of the other features. However, it appears possible to find sampling independence in some extremely simplified experimental circumstances.

5. The assumption that a given feature would have the same probability of being sampled in different letters (across-letter invariance) appeared to be viable in the present study but may not hold with all alphabets. The assumption that different features are extracted with equal overall probabilities (across-feature invariance) was substantiated, in the case of features that were not contained in all letters, by signal detectability analyses. This finding suggests that the unequal report frequencies for these features was due to decisional rather than sensitivity influences.

6. The observers reported features that were not contained in the stimulus letter (ghost features) with a significant frequency. This result falsifies the traditional high-threshold assumption of feature-detection models.

7. All the substantive decision models evidenced the proper direction of change of bias parameters with alterations in the payoff conditions. Overall, the best decision model of those tested in the present experiment seems to be the multicomponent model, accentuating as it does the importance of ghost features in rejecting alternative letter possibilities. However, there was a tendency on the part of the observers to be heavily biased toward a letter alternative perfectly matching a sampled feature set. This bias provides difficulties for the multicomponent and similar models when some letters are subsets of others, as the F (feature set 1, 3, 4) is a subset of the E (feature set 1, 3, 4, 5) in the present study. This problem might be solved in future models, although perhaps at the expense of more parameters.

8. The multicomponent decision model, when combined with the second most general feature-sampling model, predicted the individual 4 × 4 confusion matrices almost as well as did the choice model, but only after reestimation of the parameters. This model also had one more parameter than did the choice model.

The substantive feature models are more economical than are the descriptive models in that less parameter estimation should be required in other experiments using the same features; also, they yield an intuitive processing explanation of the data. In the present experiment, the feature models yielded reasonable accounts of the data on a number of dimensions, the notable exception being the failure of the assumption of sampling independence. It seems probable, from the above and other recent results, that sampling dependencies are often masked in typical studies by the added flexibility included in the decision stage of the models.

There are, finally, several studies that are

relevant to our summary discussion. Mortensen (Note 6) recently tested a feature model much like the above against a more complex feature model based on linear systems and more complex feature interactions; still, even the more complex models had difficulty with some aspects of his data.

Lupker (1979) presents evidence that is qualitatively hard to reconcile with a hierarchical feature-accumulation process where higher-order relational features depend on the accumulation of the lower-order features. His findings are qualitatively compatible with a kind of narrowing-in process where all parts of the figure are originally available but together in a blob form. As time progresses, the blob is eventually clarified, with sufficiently good contrast, to the detailed, accurate letter itself. The latter process is perhaps most readily in agreement with a form of refined template activity. Lupker's conception of the feature process contained several assumptions that must now be considered dubious, such as sampling independence, an absence of ghost features (high-threshold assumptions), or a succeeding decision stage. Although plausible arguments were offered to bolster the conclusions against hierarchical feature processing in the presence of weaker assumptions, it is difficult to know exactly which assumptions were most responsible for the falsification without tests oriented around a mathematical model. Nevertheless, his basic results are not incompatible with the present findings. For instance, if the line-segment aspects of a blob are positively or negatively correlated with one another, for a given duration of display, then sampling independence would fail in the feature reports as we found here. Further, the similarity-choice model seemed to give an adequate fit to his data, as we typically find.

Also providing some evidence against feature models was Holbrook (1975) who discovered that a simple physical overlap (template) measurement provided a better description of similarity as viewed through the choice-similarity parameters than did various feature representatives. However, that evidence is somewhat indirect. One reason is that the similarity measures within a model [illegible] on a within trial stochastic level. Thus, there is no necessary implication that such a measure on the original stimuli themselves should constitute an overall psychological "similarity" or be highly correlated with M_{ij} of the choice model. Nevertheless, it is interesting that such a simple measure of areal overlap should correlate so well with the similarity parameters of the good-fitting choice model.

In helping to refine the classes of models capable of explaining recognition and confusion data, spatial frequency representations of alphabetic characters may offer an alternative to the more obvious line-feature or visual template representations.

Coffin (1978) has suggested that spatial frequency descriptions are ineffective in representing confusability although no attempt was made to factor out bias effects. Nevertheless, sufficient success has been shown in other psychophysical and physiological contexts that models of character recognition should be constructed on the basis of spatial frequency conceptions, embedded in a probabilistic framework, and completed with a response bias process. Such models are not presently available. The simplest of such models should assume sampling independence of component frequencies. It could be that a successful model will relate at some level to the good-fitting choice model. Further, it may well be that spatial-frequency coding predominates at a lower level, whereas feature mechanisms exist at a higher level. In any case, there is no reason at present to indict the simple, data-driven aspect of the tested models as being responsible for the poor feature-sampling fits. On the other hand, cross-feed among feature channels may be an idea worth embedding in an appropriate sensory subprocess. But more empirical evidence is needed concerning the nature of feature interactions. This matter is being pursued.

When one attempts to read a degraded stimulus, such as a house number, in poor light, one's introspections suggest that discrimination between, say, an 8 and a 3 or a B and a D focus on aspects most easily described as line and curve features. Whether this kind of phenomenon will turn out to be indicative of the type of processing taking

place or rather, is an epiphenomenon associated with translation of geometric information to verbal form, will require more research.

Reference Notes

1. Gibson, E., Osser, H., Schiff, W., & Smith, J. *An analysis of critical features of letters, tested by a confusion matrix. A basic research program on reading* (Cooperative Research Project No. 639). Washington, D.C.: U.S. Office of Education, 1963.
2. Rumelhart, D. E. *A multicomponent theory of confusion among briefly exposed alphabetic characters* (CHIP 22). San Diego: University of California, Center for Human Information Processing, 1971.
3. Townsend, J. T. *Mathematical models of visual character recognition.* Paper presented at Fourth Annual Interdisciplinary Conference, Jackson Hole, Wyoming, 1979.
4. Smith, J. E. K. *Models of confusion.* Paper presented at the meeting of Psychonomic Society, St. Louis, November, 1968.
5. Hu, G. G. *Testing models of feature sampling independence.* Unpublished master's thesis, Purdue University, 1980.
6. Mortensen, U. *Models of pattern recognition by feature identification and similarity assessment* (Tech. rep.). Diskussionsbeitrage, Fachbereich Statistik, Universitat Konstanz, 1978.

References

Broadbent, D. E. Word-frequency effect and response bias. *Psychological Review*, 1967, *74*, 1–15.

Coffin, S. Spatial frequency analysis of block letters does not predict experimental confusions. *Perception & Psychophysics*, 1978, *23*, 69–74.

Falmagne, J. C. Biscalability of error matrices and all-or-none reaction time theories. *Journal of Mathematical Psychology*, 1972, *9*, 206–224.

Feigenbaum, E. A. The simulation of verbal learning behavior. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought*. New York: McGraw-Hill, 1963.

Fisher, D. F., Monty, R. A., & Glucksberg, S. Visual confusion matrices: Fact or artifact? *Journal of Psychology*, 1969, *71*, 111–125.

Garner, W. R. Aspects of a stimulus: Features, dimension, and configurations. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization*. Hillsdale, N.J.: Erlbaum, 1978.

Garner, W. R., & Haun, F. Letter identification on a function of type of perceptual limitation and type of attribute. *Journal of Experimental Psychology: Human Perception and Performance*, 1978, *4*, 199–209.

Garner, W. R., & Morton, J. Perceptual independence: Definitions, models, and experimental paradigms. *Psychological Bulletin*, 1969, *72*, 233–259.

Geyer, L. H., & DeWald, C. G. Feature lists and confusion matrices. *Perception & Psychophysics*, 1973, *14*, 471–482.

Green, D. M., & Swets, J. A. *Signal detection theory and psychophysics*. New York: Wiley, 1966.

Holbrook, M. B. A comparison of methods for measuring the interletter similarity between capital letters. *Perception & Psychophysics*, 1975, *17*, 532–536.

Hübel, D. H., & Wiesel, T. N. Receptive Fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 1962, *160*, 106–154.

Hubert, L. A statistical method for investigating the perceptual confusion among geometric configurations. *Journal of Mathematical Psychology*, 1972, *9*, 389–403.

Kuennapas, T. Visual perception of capital letters. *Scandinavian Journal of Psychology*, 1966, *7*, 189–197.

Lappin, J. S. The relativity of choice behavior and the effect of prior knowledge on the speed and accuracy of recognition. In J. Castellan & F. Restle (Eds.), *Cognitive theory* (Vol. 3). Hillsdale, N.J.: Erlbaum, 1978.

Luce, R. D. *Individual choice behavior*. New York: Wiley, 1959.

Luce, R. D. Detection and recognition. In R. D. Luce, R. B. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1). New York: Wiley, 1963.

Lupker, S. J. On the nature of perceptual information during letter perception. *Perception & Psychophysics*, 1979, *25*, 303–312.

Massaro, D. W., & Schmuller, J. Visual features, perceptual storage, and processing time in reading. In D. W. Massaro (Ed.), *Understanding language*. New York: Academic Press, 1975.

Nilsson, N. J. *Learning machines*. New York: McGraw-Hill, 1965.

Pachella, R., Smith, J. E. K., & Stanovich, K. E. Qualitative error analysis and speeded classification. In J. Castellan & F. Restle (Eds.), *Cognitive Theory* (Vol. 3). Hillsdale, N.J.: Erlbaum, 1978.

Reed, S. K. *Psychological processes in pattern recognition*. New York: Academic Press, 1973.

Rumelhart, D. E., & Siple, P. Process of recognizing tachistoscopically presented words. *Psychological Review*, 1974, *81*, 99–112.

Schulze, H. H., Baurichter, W., Gerling, W., & Grobe, R. Independent feature processing in the visual system: The effects of symmetry and distance of components. *Psychological Research*, 1977, *39*(3), 249–259.

Selfridge, O. G. Pandemonium: A paradigm for learning. In, *The mechanisation of thought processes*. H. M. Stationary Office, London, 1959.

Taylor, D. A. Holistic and analytic processes in the comparison of letters. *Perception & Psychophysics*, 1976, *20*, 187–190.

Townsend, J. T. Alphabetic confusion: A test of models for individuals. *Perception & Psychophysics*, 1971, *9*, 449–454. (a)

Townsend, J. T. Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics*, 1971, *9*, 40–50. (b)

Townsend, J. T. A stochastic theory of matching processes. *Journal of Mathematical Psychology*, 1976, *14*, 1–52.

Townsend, J. T. A clarification of some current multi-

plicative confusion models. *Journal of Mathematical Psychology*, 1978, *18*, 25–38.

Townsend, J. T., Hu, G. G., & Ashby, F. G. A test of visual feature sampling independence with orthogonal straight lines. *Bulletin of the Psychonomic Society*, 1980, *15*, 163–166.

Townsend, J. T., Hu, G. G., & Ashby, F. G. Perceptual sampling of orthogonal straight line features. *Psychological Research*, 1981, *43*, 259–275.

Townsend, J. T., & Landon, D. E. An experimental and theoretical investigation of the constant ratio rule and other models of visual letter confusion. *Journal of Mathematical Psychology*, 1982, *25*, 119–162.

Townsend, J. T., & Landon, D. E. Mathematical models of recognition and confusion in psychology. *International Journal of Mathematical Social Sciences*, in press.

Wandmacher, J. Multicomponent theory of perception: Feature extraction and response decision in visual identification. *Psychological Research*, 1976, *39*, 17–37.

Wandmacher, J. S-Multiplicativity of a stochastic matrix and applications to visual identification. *Journal of Mathematical Psychology*, 1977, *16*, 217–232.

Received October 5, 1981
Revision received June 10, 1982 ■

Instructions to Authors

For further information on content, authors should refer to the editorial in the August 1978 issue of the Journal (Vol. 5, No. 3, pp. 355–356). Authors should prepare manuscripts according to the *Publication Manual of the American Psychological Association* (2nd ed.). Instructions on tables, figures, references, metrics, and typing (all copy must be double-spaced) appear in the Manual. Manuscripts should include an abstract of 100–175 words. Authors are requested to refer to the "Guidelines for Nonsexist Language in APA Journals" (Publication Manual Change Sheet *2*, *American Psychologist*, June 1977, pp. 487–494) before submitting manuscripts to this journal.

APA policy prohibits an author from simultaneously submitting the same manuscript to two or more journals. Authors should submit manuscripts in quadruplicate, and all copies should be clear, readable, and on paper of good quality. Authors should keep a copy of their manuscript to guard against loss. Mail manuscripts to Editor-elect William Epstein, Department of Psychology, University of Wisconsin, W. J. Brogden Psychology Building, 1202 West Johnson Street, Madison, Wisconsin 53706.

Addresses for the editors of the other JEP journals are as follows: *Journal of Experimental Psychology: General*, Gregory A. Kimble, Department of Psychology, Duke University, Durham, North Carolina 27706 (submit 4 copies of the manuscript); *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Richard M. Shiffrin, Department of Psychology, Indiana University, Bloomington, Indiana 47405 (submit 4 copies of the manuscript); and *Journal of Experimental Psychology: Animal Behavior Processes*, Donald S. Blough, Department of Psychology, Brown University, Providence, Rhode Island 02912 (submit 3 copies of the manuscript).

The editors have agreed to use blind review when it is requested by the author. Authors requesting blind review should prepare manuscripts to conceal their identity.