Psychological Review
1990, Vol. 97, No. 2, 225-252

0033-295X/90/$00.75

Models of Integration Given Multiple Sources of Information

Dominic W. Massaro
University of California, Santa Cruz

Daniel Friedman
Department of Economics
University of California, Santa Cruz

Several models of information integration are developed and analyzed within the context of a prototypical pattern-recognition task. The central concerns are whether the models prescribe maximally efficient (optimal) integration and to what extent the models are psychologically valid. Evaluation, integration, and decision processes are specified for each model. Important features are whether evaluation is noisy, whether integration follows Bayes's theorem, and whether decision consists of a criterion rule or a relative goodness rule. Simulations of the models and predictions of the results by the same models are carried out to provide a measure of identifiability or the extent to which the models can be distinguished from one another. The models are also contrasted against empirical results from tasks with 2 and 4 response alternatives and with graded responses.

Conceptual Framework

There is a growing consensus that behavior reflects the influence of multiple sources of information. Auditory and visual perception, reading and speech perception, and decision making and judgment are modulated by a wide variety of influences (Anderson, 1981; Bruno & Cutting, 1988; Falmagne, 1985; Massaro, 1987a, 1988a; Oden, 1981; Perkell & Klatt, 1986; Welch & Warren, 1980). Until only recently, psychological inquiry was aimed at studying the relationship between behavior and a given single source independently of other sources of information. The common strategy was to eliminate or to hold constant all potential sources of information except the source of interest. This research strategy was most apparent in psychophysics but was also pervasive in perception, memory, and learning.

The single-factor experiment was the dominant mode of investigation when one-dimensional functional relationships were the primary goal. Trying to understand behavior when multiple sources of information are available poses additional problems. Factorial experiments seem to be the most promising approach, and we have witnessed immense methodological and theoretical progress in this domain. Specifically, the additive-factor method developed by Sternberg (1969) and Anderson's (1970, 1981) functional measurement are milestones that will not be easily surpassed. Without these methodologies, there would have been a plethora of idle psychologists in the last couple of decades. True, Fisher (1935) bequeathed the statistical tools for factorial designs long before Anderson, Sternberg, and other scientists exploited them. However, Anderson and Sternberg contributed paradigms for blending Fisher's methodology and psychological theory—something that had not been done previously. Although this blending is not without fault (Gigerenzer & Murray, 1987), the positive contributions of the research paradigms cannot be questioned (Townsend, 1984).

The magnitude of the problem of multiple sources of information compared with understanding how a single source might have an influence is uncertain. A comparable problem is illustrated by the considerable research effort that has been directed at the question of threshold versus continuous-state sensory systems and the difficulty in deciding between these two alternatives (Krantz, 1969; Massaro, 1969; Swets, 1961; Swets, Tanner, & Birdsall, 1961; Wickelgren, 1968). Given this experience, a justified fear is that the problem of how multiple sources of information influence behavior will increase in difficulty in some exponential manner. A hope is that the manipulation of multiple sources of information will also provide more experimental power than a single-factor design and, eventually, make the task easier.

In this article, we present and compare various existing models of how multiple sources of information influence perception and decision. The question we address is how individuals process two or more sources of information that may agree with one another or conflict to various degrees. The central concerns are the processes assumed by the models and resulting differences in their predictions. Our goal is to identify the similarities and differences among the models that are often overlooked in the literature. We also address the optimality properties and empirical validity of the models. Although most of our examples involve a prototypical pattern-recognition task and the application of extant models to this task, our analysis can be applied to any domain involving information integration. Each model is described and implemented, and similarities and differences among the models are noted for various types of experiments. We begin our discussion with a description of our prototypical task and a taxonomy of experiments.

Taxonomy of Experiments

We describe different types of experimental tasks and often use specific examples to facilitate the presentation. Hence, we

The writing of this article was supported, in part, by a grant from the National Institutes of Health (NINCDS Grant 20314), a grant from the National Science Foundation (BNS 8812728), and a James McKeen Cattell Fellowship to Dominic W. Massaro; and by a grant from the National Science Foundation (IRI 8812798) to Daniel Friedman.

We thank Michael M. Cohen, Yoshihisa Kashima, Roger Shepard, James Townsend, and an anonymous reviewer for their comments.

Correspondence concerning this article should be addressed to Dominic W. Massaro, Department of Psychology, University of California, Santa Cruz, California 95064.

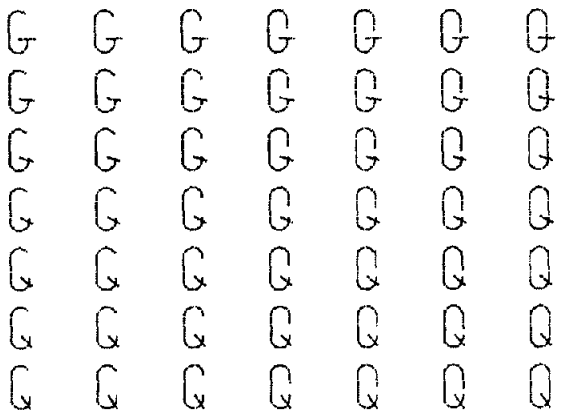

Figure 1. Forty-nine test letters, varying between *G* and *Q*, created by varying the obliqueness of the straight line (row factor) and the closedness of the gap in the oval (column factor). (After Massaro & Hary, 1986.)

begin by describing a prototypical pattern-recognition task of manipulating two sources of information at several levels. Two categories, *G* and *Q*, are chosen as the alternatives in a letter-processing task (Massaro & Hary, 1986). A factorial design is used to generate test stimuli representing all combinations of the two sources of information. A range of letters between *G* and *Q* is created when the obliqueness of a line and the closedness of the gap in the letter *Q* are varied across seven levels each (Figure 1). Seven levels of closedness are made by removing 0, 2, 3, 4, 7, 9, and 10 points from the right side of the oval of the capital letter *Q*. Similarly, the obliqueness of the line varies between the horizontal and 11, 21, 29, 38, 51, and 61 degrees of obliqueness measured from the horizontal. The resultant 49 test letters make up the factorial design. The factorial design can be expanded to allow presentation of each source of information without the presence of the other source of information. In this expanded design (not shown in Figure 1), the separate characteristics of each of the two sources of information are presented in isolation. Seven test letters are composed of just the oval, and seven test letters are composed of just the straight line. The test items are presented repeatedly to subjects in randomized order during a series of test trials. Two dependent measures are the identification judgments and the reaction times. In addition to experiments requiring categorical judgments, rating tasks can also be carried out in which subjects are asked to rate each letter along a continuum, such as that between *G* and *Q*.

We now introduce some definitions and distinctions that are useful for the developments in the article. A set of different experimental designs is illustrated in Figure 2. A *single-factor design* involves the manipulation of one independent variable. For example, only the closedness of the test letters might be varied, with the obliqueness of the straight line and all other physical properties held constant. A *factorial design* involves the orthogonal manipulation of two or more independent variables; each level of one independent variable is paired with every level of the other independent variable. In the prototypical example, this would involve using the set of 49 test letters shown in Figure 1. An *expanded factorial design* adds conditions in which each level of each independent variable is presented in isolation. The expanded conditions involve the variation of one source of information without the presence of other sources of information. In the example, the closedness of the gap in the oval of the test letter would be varied without the presence of the straight line. Analogously, the straight line would be varied without the presence of the oval. There are two types of single-factor, factorial, and expanded factorial designs. A *categorical design* involves just the endpoint stimuli of each of the independent variables. For example, the letters in the four corners in Figure 1 would make up a categorical factorial design. A *graded design* involves intermediate stimuli between the endpoints, as with the test letters in Figure 1. When the independent variables are also presented in isolation, all 63 of the test letters would constitute a graded expanded factorial design. The graded design is ideal for addressing the integration question because the exact nature of the integration can be determined only when the sources of information are varied to span the complete range of the integration function.

These experimental designs can be used with several response modes. *Categorical* responses involve a forced choice among a set of stimulus categories. In our example, categorical responses would involve identifying each test stimulus as *Q* or *G* or as one of some other set of letter categories. For example, it would not be unreasonable to give subjects the four letter alternatives corresponding to the stimuli in the corners of Figure 1. Townsend, Hu, and Kadlec (1988) suggested the term *feature complete factorial design* for an experiment using the four stimuli at the endpoints in Figure 1 along with the four corresponding response alternatives. More generally, the number of response alternatives could be as small as two or as large as the number of unique test stimuli. With respect to the test letters in Figure 1, Nosofsky (1986) and others have described the task with two response alternatives as *categorization* and the task with 49 alternatives as *identification*. For symmetrical designs that have the same number of levels of each independent variable, the number of categorical responses can be 2, or 2^n where n is the number of independent variables, or k^n where k is the number of levels of each independent variable, or in fact any value between 2 and k^n.

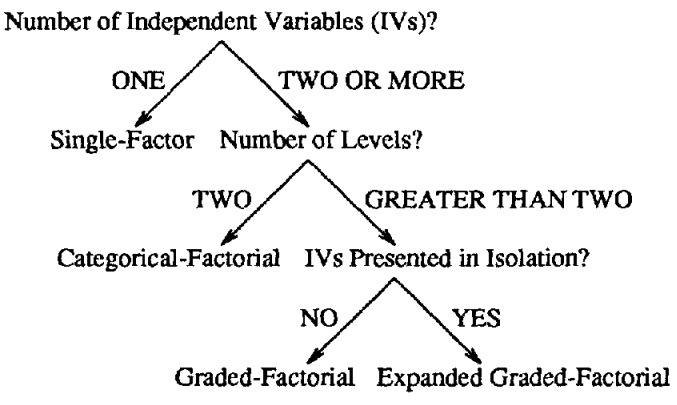

Figure 2. A taxonomy of different experimental designs illustrating some important distinctions among different types of pattern-recognition tasks.

Graded responses permit subjects to indicate the degree, probability, or confidence that the test stimulus matches one of the response categories. For example, subjects could be asked to rate on a scale from 0 to 100 the degree to which the test letter matches G as opposed to Q, where 0 is a perfect G and 100 is a perfect Q. Similarly, the subject might be given a 50-mm line between the Q and G alternatives and asked to mark the line corresponding to their interpretation of the test letter. Tasks with graded responses have been called *rating tasks* (e.g., Kornbrot, 1978). Graded-response alternatives could be multidimensional in principle, but almost all previous tasks have been limited to one dimension.

Note that pattern recognition tasks, such as the G–Q experiment, go beyond the domain typically addressed by recognition or decision models. In previous work, a normatively correct answer could be computed from the stimulus information that is given to the subject. For example, the subject is shown two urns of balls and told the percentage of red and green balls in each urn. The subject is also told the prior probability (or likelihood) that each urn of balls would be picked. Finally, the subject is told the composition of a sample of balls that was obtained in a given draw from one of the urns. The question asked is, From which urn was the observed sample taken? A normatively correct answer can be computed for this urn task, and a subject's choice can be compared with the normatively correct one. In the letter-recognition task, on the other hand, there is no objectively correct answer, and subjects are not given feedback. The subject simply gives his or her perceptual report. Even so, the graded factorial design permits us to address the question of how the information is processed and whether this information processing is optimal (see the Optimality section).

A General Stage Model

Given the prototypical pattern-recognition task, or any other psychophysical task involving multiple sources of information, a basic empirical difficulty is that information integration cannot be observed in isolation because several processes are involved. The tasks described in the previous section appear to involve evaluation, integration, and decision processes (Selfridge, 1959). *Evaluation* is defined as the analysis of each source of information by the processing system. It can be thought of as the transformation of the physical value of each source into a psychological value. In the G–Q task, for example, evaluation would give separate representations of the oval and straight-line components of the test letter. *Integration* is defined as some combination of the representations made available by the evaluation process. *Decision* maps the outcome of integration into a response. To develop the various models of pattern recognition and decision making, we give an account of these three stages of processing between stimulus and response. The three stages of processing are illustrated in Figure 3. Regardless of the type of model, each of these stages must be specified to make predictions of performance. A theory must describe how each source of information is evaluated, whether and how the different sources are integrated, and how a decision is made given the outcome of evaluation and integration.

As anticipated by Estes (1986), the models could be compared and tested more easily if our experiments could provide results about the operations of one stage without the contribution of the other stages. All three stages are not necessarily involved in all tasks, but even the simplest experiment appears to require at least two of these stages. Although integration would not occur if only one source of information were presented, evaluation of that source of information and selecting a response based on the outcome of evaluation would still be necessary. We also consider models that bypass integration and send the outputs of evaluation directly to decision. We call these models *nonintegration models*. Whether or not integration occurs, a decision process mediates the actual response. For some tasks, one might assume that the response directly reflects the outcome of integration and therefore bypasses the decision stage. This assumption has been used with considerable success in traditional psychophysical scaling (Stevens, 1961) and in information integration (Anderson, 1981).

Optimality

We make the following assumptions about the three stages of information processing. The outcome of the first stage, evaluation, can be described by a scale value, which in general we denote as x for a given information source X, y for an information source Y, and so on. The appendix is a summary of the notation used throughout this article. We assume that x is a real number on an interval scale that is measured in some sort of "currency," such as truth value, probability, activation, energy, or strength. We do not discuss binary-valued feature models; available evidence such as that presented by Shaw, Mulligan, and Stone (1983) and Massaro (1987b) suggests that real-valued evaluation functions better explain the data. For each source of information, this scale value is some function (possibly stochastic) of the stimulus provided from that source but is independent of the stimulus from other sources. In the stochastic case, this assumption naturally can be extended to perceptual independence in the sense described by Ashby and Townsend (1986). For example, in the G–Q recognition experiment, our independence assumption rules out the possibility that an observer's evaluation of the degree of closedness is affected by the level of obliqueness in a test object. However, it certainly does not rule out a statistical interaction of the sources in his or her responses that is due to the nature of the integration process.

We assume that the information-integration stage, our central concern, provides a single scale value a_k (measured in the same currency as the x and y) as a deterministic function of the scale values provided by the evaluation stage, for each choice alternative A_k. At this point, we put no restrictions on the functional form of the integration function, so as to allow investigation of a wide class of integration models. Although we will not emphasize them in this article (see the First-Order Versus Second-Order Integration section for a brief discussion), this formulation includes so-called nonintegration models. We emphasize that the value a_k is assumed to have no "memory" of how it was obtained. If two different combinations of the sources of information lead to identical outcomes of integration, then the decision would be the same in both cases. Put in somewhat different terms, the decision process does not have access to the initial statistics given by evaluation and operates on only the summary statistic produced by integration. In the case of two

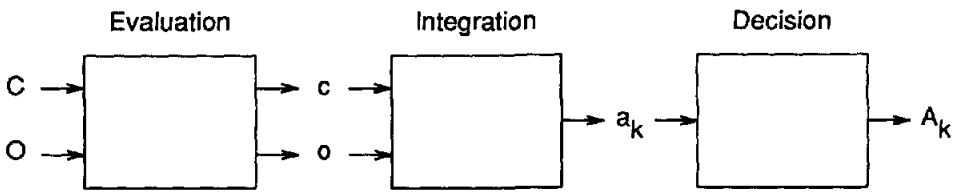

Figure 3. Schematic representation of the three stages of processing that must be accounted for in a pattern recognition-task with multiple sources of information. (The three stages are illustrated for the sources of information closedness C and obliqueness O in the G–Q task. The evaluation of the degree to which the oval is closed and the straight line is oblique produces values c and o that are made available to the integration process. If no integration occurs, these two values are passed directly to decision. Integration of the values gives an overall value a_k, indicating the degree of support for alternative k [A_k]. The decision process maps the information made available to it into a response.)

response categories or a graded response between two categories, a single value a can represent the outcome of integration (and the subscript can be dropped).

Regarding the final stage, decision, we assume only that the value of the chosen response alternative A_k is some deterministic or stochastic function of the integrated scale values a_k of all relevant alternatives. Two of the more popular decision rules, which we refer to as the *criterion rule* (CR) and the *relative goodness rule* (RGR), are described in the Ingredients for Integration Models section.

In our examination of models from the literature, we specify the evaluation and decision stages as well as the integration stage, although one or more of these stages is often left implicit in the original presentation. An important concept to recognize is that the validity and optimality of a given model of information integration generally depend on its assumptions regarding the evaluation and decision stages, as well as its specification of the integration stage (Estes, 1986).

We are now prepared to define optimality as a property of an integration model. The basic idea is that the integration process is optimal if it maximizes the final information content or, equivalently, minimizes the average loss of information. Specifically, if a_k is a sufficient statistic (DeGroot, 1970) for (x, y, $\cdots$), then there is no loss of information in the integration stage, and the integration function is optimal. Often, no sufficient statistic exists, and optimality must be judged in terms of all three stages taken together. In this case, we use the usual definition from statistical decision theory: For a given reward structure for responses and given structure for presenting stimuli (possibly including noise), the overall process is optimal if it maximizes the expected reward. If the reward structure and stimuli presentation are unbiased (in senses to be discussed below), then this weaker notion of optimality reduces to the maximum likelihood property: An individual chooses the response that has the greatest likelihood of being correct.

For example, if the currency (i.e., the scale values produced by evaluation and used by the decision process) is subjective probability, then Bayes's theorem, discussed in the next section, always produces a posterior probability that correctly and fully incorporates the prior probabilities and likelihoods obtained from the evaluation stage. Hence, this posterior probability is a sufficient statistic, and the integration process that produces it is optimal. We show that some models with currency that is not subjective probability also produce sufficient statistics in some contexts. However, for most models, we investigate optimality of the overall prediction (evaluation and decision together with integration), usually with reference to maximum likelihood. We emphasize that the *sufficient-statistic* definition of optimality allows subjective probabilities used by the subject to differ from objective probabilities. In many tasks, in fact, objective probabilities do not exist (see the Taxonomy of Experiments section). When they do, an optimal integration process might not maximize the objective expected value.

Note that optimality differs from empirical validity. Independently of the optimality question, we also ask to what extent a given model accurately describes the actual results of an experiment (see the Empirical Predictions and Tests of the Models section). This analysis of empirical results also addresses the interesting question of whether human choice behavior is optimal.

Implementation of Models

We illustrate model implementation with results drawn from Massaro and Hary (1986), who actually carried out the letter-recognition task that we have described, using a graded factorial design. Nine subjects saw each of the test letters (shown in Figure 1) for 400 ms 12 times in random order. On each trial, they labeled the test letter Q or G. Figure 4 gives the observed performance for 2 subjects. The probability of a Q response for each test letter is the dependent variable. Given that the Q and G identifications sum to 1, the probability of a Q response to each test letter, $P(A_Q)$ completely represents the identification judgments. Thus, we have 49 independent observations to describe the 49 test letters.

The ultimate goal of our analysis of integration models is to determine their optimality properties and to discover which models better describe actual behavior. It is important to keep in mind how each of the models is implemented in a given experiment. All of the models require free parameters. That is, none of the models specifies a priori the outcome of evaluation for a given level of a given source of information. However, the models should have equivalent degrees of freedom when confronting our basic pattern-recognition task so a valid comparison can be made. We limit the number of free parameters to the number of unique levels of the independent variables. In the G–

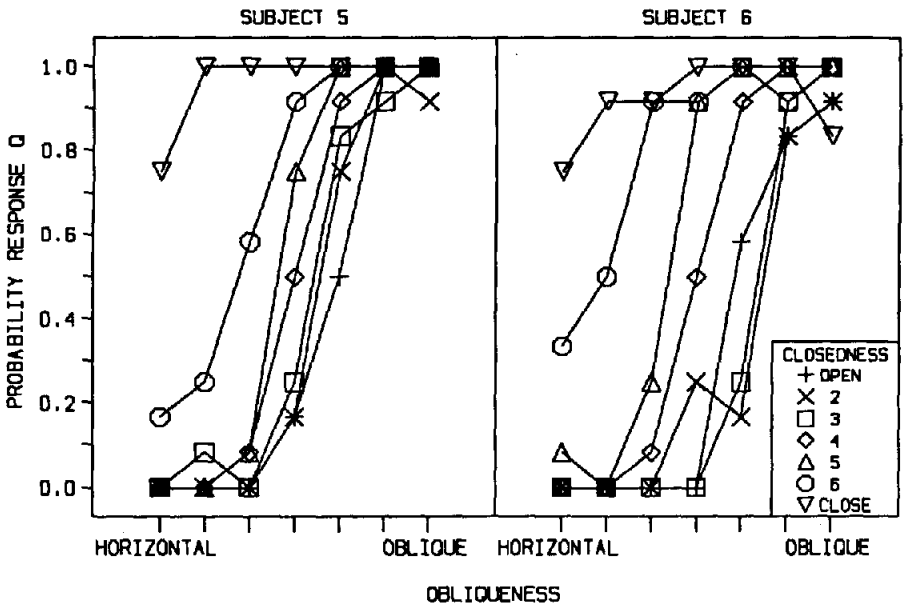

Figure 4. Observed probability of *Q* responses for the 49 test letters presented in Figure 1 (created by varying the obliqueness of the straight line and the closedness of the gap in the oval). (The results are for 2 typical subjects from an experiment carried out by Massaro and Hary, 1986.)

Q task illustrated in Figures 1 and 4 (or in the expanded version of this task), 14 parameters are necessary: seven parameters each for obliqueness and closedness. The parameters represent some measure of the degree to which the obliqueness and closedness features are present in the test letter. These estimated parameters in turn give rise to specific predictions for each model regarding response frequencies (or response ratings in a graded design). We contrast these predictions in the Empirical Predictions and Tests of the Models section.

Ingredients for Integration Models

We postulate evaluation, integration, and decision processes and illustrate the importance of each of these contributions in our analyses of the models. In this section, we make some preliminary remarks addressing the issue of how each of these processes influences optimality. Assumptions about evaluation have consequences for optimality. The primary consideration is whether evaluation is noisy (stochastic) or noise free (deterministic). To anticipate, most of the models we consider assume noise-free evaluation, whereas the models based on statistical theory, such as the theory of signal detectability (TSD), typically assume a noisy evaluation process. Assuming an independent sample of noise added to the evaluation of each source of information usually leads to different predictions than assuming noise-free evaluation with noise added at some later stage, for example, in a stochastic decision process.

Bayes's Theorem

The most venerable method for combining multiple sources of information is given by a theorem attributed to Reverend Thomas Bayes (circa 1701–1761) but also derived independently by Pierre Laplace (1749–1827; Stigler, 1986). Bayes's theorem states that

$$P(H_1|E) = \frac{P(E|H_1) \times P(H_1)}{\sum_i P(E|H_i) \times P(H_i)}, \quad (1)$$

where $P(H_i|E)$ is the probability that some hypothesis H_i is true given that some evidence E is observed; $P(E|H_i)$ is the probability of the evidence E, given that the hypothesis H_i is true, and $P(H_i)$ is the a priori probability of the hypothesis H_i. The probability of hypothesis H_1 given some evidence E is equal to the probability of the evidence given the hypothesis times the a priori probability of the hypothesis, divided by the sum of analogous likelihoods for all possible hypotheses. If the a priori probabilities of all possible hypotheses are equal, Bayes's theorem reduces to

$$P(H_1|E) = \frac{P(E|H_1)}{\sum_i P(E|H_i)}. \quad (2)$$

Bayes's theorem specifies how different sources of evidence are combined. Given two independent pieces of evidence E_1 and E_2 and equal a priori probabilities, the probability of a hypothesis H_1 is equal to

$$P(H_1|E_1 \text{ and } E_2) = \frac{P(E_1 \text{ and } E_2|H_1)}{\sum_i P(E_1 \text{ and } E_2|H_i)} = \frac{P(E_1|H_1) \times P(E_2|H_1)}{\sum_i P(E_1|H_i) \times P(E_2|H_i)}. \quad (3)$$

Equation 3 has a direct correspondence to our evaluation and

integration processes. In our notation, Equation 3 gives the outcome a_1 for integrating two sources of information X and Y, where $P(E_1 | H_1)$ represents evaluation of the first source (in terms of the subjective probability currency) and $P(E_2 | H_1)$ represents separate evaluation of the second source. Equation 3 describes optimal information integration in the currency of probability under two assumptions. First, the prior probabilities of all relevant response alternatives are equal. Second, the sources of evidence are evaluated independently of one another, as explained previously in the Optimality section. Under these assumptions, Equation 3 follows from probability theory, in which the probability of the joint occurrence of two independent events is the multiplicative combination of the probabilities of the separate events. The probability of two heads in two tosses of a coin, for example, is the multiplicative combination of the probability of a head on each toss. See Stigler (1986) for the derivation, which of course goes back 200 years to Reverend Bayes and Laplace.

Criterion Rule (CR)

As illustrated in Figure 3, the outcome of integration is transformed by a decision process to produce a response. We consider two general algorithms for the decision operation. The first, derived from communication theory, rests on the notion of a criterion. The decision operation uses a criterion value to assess the outcome of integration (or evaluation in the case of a single source of information). In a task with two response alternatives, for example, the outcome is compared with the criterion. If the outcome exceeds the criterion, one of the alternatives is selected. Otherwise, the other alternative is selected.

Consider a stimulus continuum in a graded single-factor design in which the value of information source X is varied from *not A* to A. Assume, for this argument, that this variation gives linearly increasing evidence for a given alternative A. That is, the outcome of evaluation (or integration, given multiple sources of information) is assumed to be a linear function of some independent variable. The left panel of Figure 5 shows this outcome as a linear function of variable X.

A deterministic criterion rule in a discrete judgment task with the criterion value at .5 would classify the pattern as A for any value of a greater than this criterion value. Otherwise, the pattern is classified as *not A*. Given this CR, the probability of an A response would take the step-function form shown in the right panel of Figure 5. That is, with a fixed criterion value and no noise, the decision operation changes the continuous linear function of a into a step function of *probability of response* (A). Although based on continuous evidence, the response function is discrete. This categorical result is uncommon for actual experiments (see Figure 4).

If there is noise in the mapping from variable X to a, however, a given level of variable X cannot be expected to produce the same identification judgment on each presentation. With the addition of noise, it is reasonable to assume that a given level of variable X produces a bell-shaped range of values of a with a mean directly related to the level of variable X and a variance equal across all levels of variable X. Figure 6 illustrates the expected outcome for identification if there is bell-shaped noise added to a with the same criterion value assumed in Figure 5. A signal with a mean value of a at the criterion value will produce completely random classifications over many trials. This value of a based on both signal and noise is above the criterion on half of the trials and below the criterion on the other half. As the mean of variable X moves away from the criterion value, the addition of noise will have a diminishing effect on the identification judgments. Thus, noise will have a larger influence on identification in the middle of the range of probability values than it will at the extremes. A similar outcome to that shown in Figure 6 is achieved if the mapping from variable X to a is noise free and the criterion value fluctuates randomly from moment to moment (Carterette, Friedman, & Wyman, 1966).

Relative Goodness Rule (RGR)

A second algorithm for decision is based on the ideas of Shepard (1957, 1986), Clarke (1957), Luce (1959, 1977), and Anderson (1981). This is the RGR algorithm. Two underlying assumptions are that alternatives defined as irrelevant to the choice task play no role in the decision and that the probability of a response alternative is simply equal to the ratio of the goodness of match of that alternative relative to the sum of the goodness of matches of all relevant alternatives. In the context of a categorical-response experiment with m alternatives, this general rule can be expressed as

$$P(A_k) = \frac{a_k}{\sum_{i=1}^{m} a_i}, \tag{4}$$

or the expected probability of response A_k is equal to the scale value a_k of that alternative divided by the sum of the scale values for all the relevant alternatives in the task (including the alternative of interest). In contrast to the deterministic algorithm based on a CR, the RGR predicts a response only probabilistically. The RGR specifies only asymptotic response probabilities; it is not a complete process model of how these probabilities occur. Townsend and Landon (1982) provide a few alternative process interpretations that are consistent with the choice rule, but there have been no tests among these alternatives. Although we lack a process model, there is considerable evidence that judgment appears to be relative, as predicted by the RGR (Luce, 1977; Oden, 1977).

Applying the RGR when the currency is subjective probability creates a situation called *probability matching* (Davison & McCarthy, 1988; Thomas & Legge, 1970). That is, subjects might not respond optimally by always choosing the most likely alternative but might instead choose each alternative with the probability given by Bayes's theorem. This model predicts that the probability of a response corresponding to hypothesis H_i is given by Equation 3. Although nonoptimal, this prediction should be taken seriously, given that humans and animals have been shown to probability match in many different domains (Davison & McCarthy, 1988; Estes, 1984; Myers, 1976).

In experiments with graded responses, the RGR is straightforward. For example, in our prototypical pattern-recognition task, Equation 4 would apply to continuous rating judgments on individual trials, not just average probability of categorical-response alternatives. In contrast to the RGR prediction for cate-

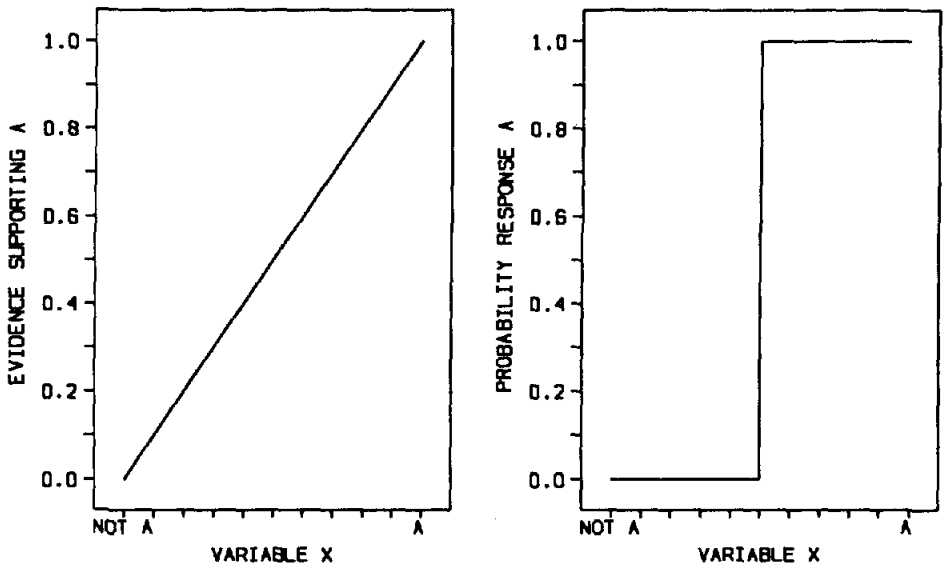

Figure 5. Left panel: The evidence for A as a function of the level along a stimulus continuum between *not* A and A. Right panel: The probability of an A response as a function of the stimulus continuum if the subject maintains a criterion rule at a particular value and responds A if and only if evidence for A exceeds the criterion.

gorical responses, its predictions for graded responses are optimal as long as the response on each trial can be interpreted as a subjective probability. Given the optimality of RGR for graded responses, an argument might be made for optimality of RGR for categorical responses. In this case, the decision maker's goal is to communicate subjective probability over the course of the experiment rather than simply the most likely alternative on any given trial. We accept this logic in our analysis.

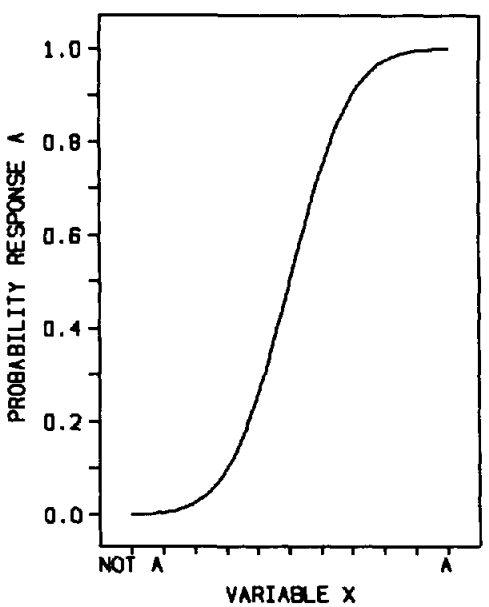

Figure 6. The probability of an A response as a function of variable X given the linear relationship between evidence for A and variable X and the criterion rule represented in Figure 5, but with bell-shaped (truncated normal) noise added to the mapping of variable X to evidence for A.

Fuzzy-Logical Model of Perception (FLMP)

We begin our survey of specific models with the fuzzy-logical model of perception (FLMP) for several reasons. First, the research framework we used for this article emerged together with the model over the course of empirical and theoretical work. Second, the model, although developed independently of Bayes's theorem, has identical optimality properties for integration. Third, the three operations of evaluation, integration, and decision are clearly articulated in the model.

Underlying this model is the assumption that well-learned patterns are recognized in accordance with a general algorithm, regardless of the modality or particular nature of the patterns (Massaro, 1984, 1987a; Oden, 1981, 1984). The model has received support in a wide variety of domains. The model consists of three operations in perceptual recognition: feature evaluation, feature integration, and pattern classification. Continuously valued features are evaluated, integrated, and matched against prototype descriptions in memory, and an identification decision is made on the basis of the relative goodness of match of the stimulus information with the relevant prototype descriptions.

Given multiple features, it is useful to have a common metric representing the degree of match of each feature. Two features that define a prototype can be related to one another more easily if they share a common currency. To serve this purpose, fuzzy-truth values (Goguen, 1969; Zadeh, 1965) are used because they provide a natural representation of the degree of match. Fuzzy-truth values lie between 0 and 1, corresponding to a

proposition being *completely false* and *completely true*. The value .5 corresponds to a completely ambiguous situation, whereas .7 would be more true than false and so on. Fuzzy-truth values, therefore, not only can represent continuous rather than just categorical information, they also can represent different kinds of information.

The three operations between presentation of a pattern and its categorization, as illustrated in Figure 3, can be formalized mathematically. Feature evaluation gives the degree to which a given dimension supports each test alternative. The physical input is transformed to a psychological value and is represented in lowercase letters; for example, dimension X would be transformed to x_k, and analogously for dimension Y. Each dimension provides a feature value at feature evaluation. Feature integration consists of a multiplicative combination of feature values supporting a given alternative A_k. If x_k and y_k are the values supporting alternative A_k, then the total support a_k for the alternative A_k would be given by the product of x_k and y_k.

The third operation is pattern classification, which gives the relative degree of support (merit) for each of the test alternatives. In this case, the probability of response A_k given the specific stimulus $X_i Y_j$ is

$$P(A_k | X = X_i, Y = Y_j) = \frac{x_k y_k}{\sum_{i=1}^{m} x_i y_i}, \tag{5}$$

where the denominator is equal to the sum of the merit of all m relevant alternatives, derived in the same manner as illustrated for alternative A_k.

To recapitulate, evaluation in the FLMP involves the representation of each source of information in terms of a truth value, between 0 and 1, indicating the merit of a particular alternative. Integration consists of a multiplicative combination of truth values. The decision uses the RGR.

Implementation of FLMP

Given a test letter in the G–Q task, the featural evaluation stage determines the degree to which the Q and G alternatives are supported by each feature of the visual information. With the use of fuzzy-truth values, a value between 0 and 1 is assigned to the oval and straight-line dimensions, indicating the degree to which these features support the Q and G alternatives. These feature values are then integrated within the Q and G prototypes. The prototypes are defined by:

Q: closed oval and oblique line

and

G: open oval and horizontal line.

Given a prototype's independent specifications for the oval and straight-line features, the value of one of these features cannot change the value of the other feature at feature integration. In the implementation of the model, *closed* and *open* are assumed to be opposites (or negations) of one another, as are *oblique* and *horizontal*. Using the definition of *fuzzy negation* as 1 minus the feature value (Zadeh, 1965), we can represent the prototypes in terms of the degree to which the oval is closed and the line is oblique:

Q: closed and oblique

and

G: (1 − closed) and (1 − oblique).

The integration of the features defining each prototype can be represented by the product of the feature values (Oden, 1979; Oden & Massaro, 1978). In this case, the goodness of match with a Q or G alternative can be represented by

$$a_Q = c \times o$$

and

$$a_G = (1 - c) \times (1 - o),$$

where a_Q and a_G represent the goodness of match of a test letter to the Q and G alternatives, respectively.

If Q and G are the only valid response alternatives, the decision operation determines their relative merit, leading to the prediction

$$P(A_Q) = \frac{a_Q}{a_Q + a_G}, \tag{6}$$

where $P(A_Q)$ is the predicted probability of a Q response to a particular test letter shown in Figure 1. In graded-response tasks, Equation 6 gives the mean predicted rating linearly scaled between 0 and 1.

Comparison of FLMP and Bayesian Integration

The FLMP is closely related to Bayesian integration. The concept of fuzzy-truth value differs from that of a subjective conditional probability (see the Previous Rejections of Optimal Behavior section). However, if the two concepts are assumed to coincide for a particular prediction, then simple substitution shows that Equations 3 and 5 are identical. That is, Bayes's theorem and the FLMP are conceptually equivalent if the truth value can be interpreted as a conditional probability.

Even if truth values and probabilities are conceptually different mappings of evaluated information from a single source into scale values, their estimated values in empirical tests will be the same. For instance, in the case of an expanded factorial design with categorical responses for our prototypical pattern-recognition task, the number of parameters and their estimated values are the same whether they are called subjective probabilities or truth values. Therefore, the FLMP is observationally equivalent to Bayesian integration.

Theory of Signal Detectability (TSD)

A second model of combining evidence from multiple sources is derived from Thurstone's (1927) law of comparative judgment. In Case V of Thurstone's theory, the discriminal process corresponding to an object in a set of objects can be represented by a scale value that is a constant plus an independent normally distributed variable. The scale values differ across objects, but the random variable is identically distributed. Stimuli

$X_1, X_2, \ldots, X_n$ are represented psychologically by real-valued random variables $x_1, x_2, \ldots, x_n$, called *discriminal processes*. Given two response alternatives A_i and A_j, corresponding to X_i and X_j, the subject chooses A_i if and only if $x_i > x_j$. Given this assumption, the probability that A_i will be chosen is

$$P(A_i|X_i, X_j) = P[x_i > x_j] = P[x_i - x_j > 0]. \quad (7)$$

By assuming that the real-valued random variables are normally distributed with equal variance, the probability values can be transformed into scale values of the discriminal processes (d' values in the TSD). (This transformation is exactly that used by Dosher, Sperling, & Wurst, 1986, and by Bruno & Cutting, 1988, in their analyses of factorial experiments.) Thurstone's Case V then becomes mathematically equivalent to the TSD with two test stimulus alternatives, and our analyses are made on this version of the theory.

The traditional assumption in psychophysics since the time of Fechner (1801–1887) is that sensory systems are characterized by *thresholds*. A threshold represents a barrier in the sensory system that must be overcome in order for a signal to be detected. All inputs below the threshold value go undetected and have no differential influence on the sensory system. Input values above the threshold value are detected. The theory of signal detectability denied the presence of a threshold and claimed that some sensory information is always available to the sensory system (Tanner & Swets, 1954). Detection of a stimulus is viewed as being analogous to a statistical decision task in which the decision system assigns conditional probabilities to the output of the sensory system. The decision system supposedly knows the potential outputs from the environmental events of interest (as it does in a Bayesian analysis). Consider the standard signal-detection task in which there are two types of trials: noise (N) trials and signal-plus-noise (SN) trials. The decision system has knowledge of the SN and N distributions and, given evaluated stimulus x, computes the conditional probability that x arose from an N trial and the probability that it arose from an SN trial. The decision system computes a likelihood ratio equal to the probability that x occurred given SN divided by the probability that x occurred given N:

$$l(x) = \frac{P(x|SN)}{P(x|N)}. \quad (8)$$

The decision system establishes a criterion value, and if the likelihood ratio given by Equation 8 exceeds this value, the observer responds yes; otherwise, the response is no.

In this traditional signal-detection task, a measure of sensory performance that is independent of the criterion value that was used in the task can be computed. The values of x from a particular type of trial (SN or N) are assumed to be normally distributed. In addition, the variance from SN trials is usually assumed to be equal to the variance from N trials. If the scale is chosen so that the variance is equal to 1, then the distances along the x axis can be expressed in z scores. The distance in z-score units between the mean of the SN distribution and the criterion value can be computed from the hit rate $P(Yes|SN)$, and the distance between the mean of the N distribution and the criterion value can be computed from the false-alarm rate $P(Yes|N)$. The sum of these two distances preserving the sign gives d', the distance between the means of the two distributions.

Our goal, of course, is to develop the signal-detection model to address the problem of integrating multiple sources of information. An early application was Green and Swets's (1966) investigation of the relationship between yes–no tasks, in which the subject has only one observation interval before making a decision, and two-interval forced-choice tasks, in which the subject has two observation intervals before making a decision. Green and Swets (1966) assumed that the subject integrates the information by simply adding the evaluation outputs from the two observation intervals and responds on the basis on this sum. This new observation has more information relative to the single-observation condition which leads to a larger d' value. They proved that for an optimal observer, the d' value determined from two observation intervals should be the square root of 2, or 1.414 times the d' value determined from a single observation interval (Green & Swets, 1966, Appendix 9-A). This optimality result can be explained intuitively by recalling the well-known statistical result that the mean (together with the sample size) of an independent random sample drawn from a normally distributed population is a sufficient statistic for the sample. Thus, the sum $x + y$, together with the sample size of 2, carries the same information as the original sample $\{x, y\}$, and the TSD model ensures that it is properly processed in this context.

Generalizing this derivation for two observations of the same source to a single observation of two sources of information, Green and Swets (1966) stated that the d' given two sources of information, say X and Y, is equal to the square root of the sum of squares of each of the individual d' values:

$$d'_{XY} = \sqrt{(d'_X)^2 + (d'_Y)^2}. \quad (9)$$

Underlying this formula is the assumption that the observer knows the precision of each information source and takes a weighted sum of the evaluation outputs, with greater weight on the more precise (i.e., lower noise variance) source. They proved that Equation 9 is consistent with statistical (i.e., Bayesian) decision theory and therefore optimal (for an appropriate decision rule, e.g., the CR) under the following assumptions: (a) All stimuli are degraded by random noise; the output from evaluating stimulus X_i can be represented by the real number $x_i = s_i + e_i$, where e_i is a random-error term (arising from imperfect presentation or imperfect evaluation of the stimulus or both) and s_i is the evaluation of X_i in the absence of such noise; (b) the errors e_i are independent and have a mean of 0; (c) the errors e_i are normally distributed; and (d) the errors e_i have the same variance for every level of each information source but may have different variances for different information sources (Peterson, Birdsall, & Fox, 1954).

To summarize, the main assumption of the TSD model is that evaluation is degraded by noise and produces a normally distributed scale value for each source. It transforms these scale values by the inverse cumulative-unit normal distribution (z transformation) into d' values. The integration function is defined on these d' values by Equation 9. The decision process uses the CR.

Implementation of TSD: Expanded Nonfactorial Design

A natural implementation of TSD is the case of a nonfactorial expanded design with two categorical responses (Stanislaw, 1988). It is straightforward because, in contrast to the other designs, a measure of response accuracy can be defined. For example, in the prototypical pattern-recognition task shown in Figure 1, the test letters would be the upper left and lower right letters, corresponding to a prototypical G and Q, respectively. That is, the test letters in this case would be either $X_G = (C_1, O_1)$—that is, the not-closed oval with a 10-point gap and a not-oblique line, as in the letter G—or else $X_Q = (C_7, O_7)$—that is, a closed oval and a 61° oblique line, as in the letter Q. In addition, each of the two levels of the two sources of information would be presented in isolation. That is, only the closed oval, open oval, oblique line, or horizontal line is presented on these single-source trials. The allowable responses for both types of trials would be either A_Q or A_G, meaning, "it is most consistent with a Q" or "it is most consistent with a G." Following Equation 9, performance given both sources of information is predicted from performance given each of the two sources presented alone. One calculates d'_C from the relative frequencies of hits $p = P(A_Q|C_7)$ and false alarms $q = P(A_Q|C_1)$ for the oval stimulus presented in isolation by the standard formula, that is, $d'_C = Z(p) - Z(q)$, where $Z(\bullet)$ is the z-score or inverse-unit normal cumulative distribution function (CDF). (Presumably, $p > q$, so d' is positive; that is, a Q-like stimulus is more likely than a G-like stimulus to generate a Q response.) Similarly, one calculates d'_O from the Z scores for $P(A_Q|O_7)$ and $P(A_Q|O_1)$, the relative frequencies of Q responses given Q-like and G-like line stimuli in isolation. The TSD model then predicts that d'_{CO}, the d' value obtained as the difference of the Z scores for hit and false-alarm rates $P(A_Q|X_Q)$ and $P(A_Q|X_G)$ for the combined stimuli, will result from Equation 9.

Optimality in this implementation of the TSD model requires the four assumptions listed above, and these assumptions might not hold in the prototypical pattern-recognition task. If there is noise at evaluation, the noise from one source may be perfectly correlated with the other, contrary to Assumption b. In this case, as noted by Fidell (1970), Equation 9 must be replaced by the simple summation of the separate d's. Likewise, if the noise processes are not precisely normal, then no weighted sum of the evaluation outputs is a sufficient statistic, and the optimality argument fails. One can construct an example with approximately normal noise (provided by two dice) that shows dramatic failure of optimality. Finally, suppose that Assumptions a, b, and c hold, but the noise variance for the oval stimulus is slightly different for C_1 than for C_7. Then it is easy to see that the likelihood ratio $l(x_C)$ is no longer a monotonic function. In this case, the criterion rule no longer represents an optimal decision process (Green & Swets, 1966). Thus, even in its natural implementation, the TSD model becomes nonoptimal with violations of its apparently minor assumptions.

Implementation of TSD: Graded Factorial Designs

Some additional assumptions have to be made to apply the TSD model to graded factorial designs because in many cases, there is no correct answer. In recognizing uppercase letters, for example, it is not obvious which letters in Figure 1 should be called G or Q. In fact, one goal of the experiment is to determine how the subject classifies a pattern varying with respect to these levels of information. Thus, we are obtaining a perceptual report on the part of the subject that might be used to describe the relationship between the stimulus information and the perceptual judgment (see Braida & Durlach, 1972). The measure of performance now provides a measure of the consistency in categorizing stimuli, rather than the subject's reliability in distinguishing signal from noise. That is, two stimuli are considered to be highly discriminable from one another if they are consistently categorized as different stimuli (i.e., produce different responses).

Consider a response to a single dimension of the stimulus, for example O_j. The probability of a Q response given stimulus O_j, $P(A_Q|O_j)$, can be expressed in discrimination units. In this case, the subject needs to have some representation of each of the response patterns relevant to the task at hand. That is, the subject is assumed to have information in memory about the uppercase letters Q and G. A test letter is evaluated in terms of the degree to which it matches the prototypical patterns stored in memory. Or, equivalently, taking the signal-detection perspective illustrated in Figure 7, the subject can be assumed to evaluate the test letter along a one-dimensional G–Q continuum of information. The subject is assumed to place the criterion at a point equidistant between the means of the distributions corresponding to the prototypical G and Q, respectively. In our example, presentation of a given pattern produces a certain amount of Q-ness, and the subject decides whether this amount of Q-ness exceeds the criterion value separating the G and Q categories. If the observation exceeds this criterion value, the subject responds Q; otherwise, the subject responds G. Of course, the evaluation process is degraded by normally distributed noise as assumed in the standard signal-detection model.

Given this conceptualization of the task, the distance between the mean of any distribution and the criterion can be measured. For argument's sake, assume that the distribution is normal with the same variance as the prototypical distributions, so distance is measured in z scores. To the extent that this distribution is far from the criterion, subjects would show good discrimination. In this case, a given pattern would tend to be identified most of the time as G or most of the time as Q. Poor discrimination would be reflected by a small distance between the distribution and the criterion, with the subject about equally likely to identify a given pattern as G or Q. Performance is evaluated in terms of the degree to which a given stimulus pattern leads to consistent or inconsistent responses. The probability of a Q response given pattern O_j can be considered to be the hit rate, and the probability of a G response given pattern O_j can be defined to be equal to the false-alarm rate. Given that only one stimulus pattern was presented, we see that the false-alarm rate must be equal to 1 minus the hit rate. Given these hit and false-alarm rates, the distance between the mean and the criterion can be computed in the standard manner.

The TSD model represents integration by a sum of evaluations, $x_C + y_O$, with the corresponding d'_{CO} given by Equation 9. Given this model, d'_{CO}, corresponding to performance given closed and oblique characteristics of the test letter, is equal to the square root of the sum of the squared d'_C, given the closed

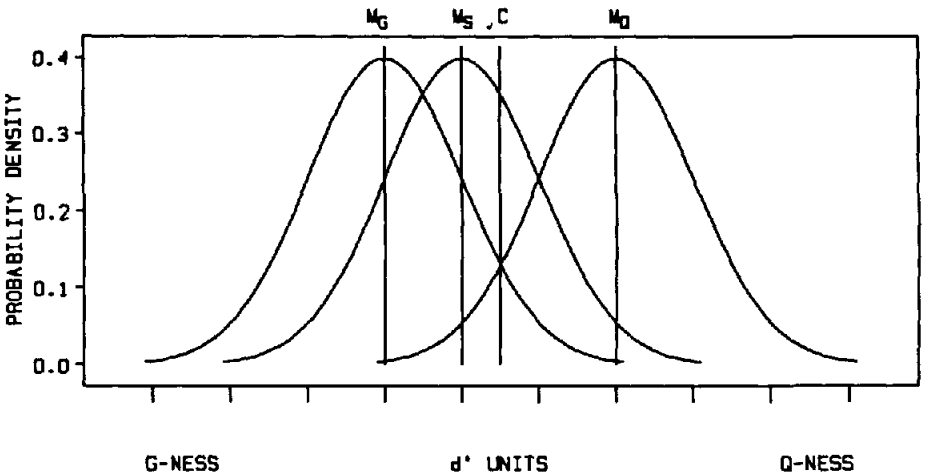

Figure 7. Two distributions corresponding to the G and Q prototypes, with a criterion C placed equidistant between the means M_G and M_Q of the G and Q distributions. (A third distribution corresponds to a test stimulus S, with mean M_S.)

characteristic, and the squared d'_O, given the oblique characteristic. (Note that the sign has to be preserved in actual practice, as indicated later in Equation 11. That is, the closed characteristic might favor G, and the oblique characteristic might favor Q, for example, and the positive d' for the oblique characteristic would be offset by the negative d' for the closed characteristic. In this case, their combination would produce relatively random judgments.)

This extension of the TSD model differs from the FLMP model. To see this, consider any expanded factorial design for the prototypical pattern-recognition task illustrated in Figure 1. Given the test alternatives Q and G, for example, the subject's response to stimulus C_iO_j on a single graded-response trial (or average response in a series of categorical-response trials) may be interpreted as the subjective probability that the stimulus was a Q, that is, $P(Q|C_iO_j)$. The corresponding responses to the single-factor trial in this case represent $P(Q|C_i)$ and $P(Q|O_j)$, respectively. According to the FLMP, the prediction is derived from

$$P(Q|C_iO_j) = \frac{P(Q|C_i) \times P(Q|O_j)}{P(Q|C_i) \times P(Q|O_j) + [1 - P(Q|C_i)] \times [1 - P(Q|O_j)]}. \quad (10)$$

Our implementation of TSD suggests a different prediction. Given that the hit rate is 1 minus the false-alarm rate, we calculate $d'_{C_i} = 2 \times Z[P(Q|C_i)]$ and $d'_{O_j} = 2 \times Z[P(Q|O_j)]$ from single-factor trials. The prediction, then, is that the $d'_{C_iO_j}$ observed from the factorial trial, calculated as $d'_{C_iO_j} = 2 \times Z[P(Q|C_iO_j)]$, will be equal to $\pm\sqrt{|(d'_{C_i})^2 \pm (d'_{O_j})^2|}$, where the minus sign of the plus/minus sign within the radical applies if d'_{C_i} and d'_{O_j} differ in sign, and the minus sign of the plus/minus sign outside the radical applies if the quantity within the $\sqrt{|\ |}$ is negative. Using $N(\bullet)$ to denote the normal CDF, the inverse of which is $Z(\bullet)$, we can reexpress this prediction as

$$P(Q|C_iO_j) = N(\pm\tfrac{1}{2}\sqrt{|(d'_{C_i})^2 \pm (d'_{O_j})^2|}). \quad (11)$$

Clearly, the predictions in Equations 10 and 11 are quite different. For example, suppose $P(Q|C_i) = P(Q|O_j) = .7$. Then, $d'_{C_i} = d'_{O_j} = 2 \cdot Z(.7) = 1.05$, and Equation 11 yields an integrated d' of $(\sqrt{2})(1.05) = 1.485$, with corresponding probability $N(\tfrac{1}{2} \cdot 1.485) = .7711$. By contrast, Equation 10 yields integrated probability $(.7)^2/[(.7)^2 + (.3)^2] = .8448$, with a corresponding d' of 2.028. Thus, this TSD model and the FLMP give different predictions for the integration of two sources of information.

To summarize, the TSD model applies directly to nonfactorial designs with correct answers and can be extended to graded factorial designs in experiments with two response alternatives. It assumes that evaluation is degraded by noise, but the sensory output x always generates one response if it exceeds some specific criterion value and otherwise always generates the alternative response. Integration in the TSD model occurs by summing the x values obtained from the independent sources of information. As stated by Green and Swets (1966, p. 271), "The so-called integration model associated with detection theory assumes in each instance that the multiple observations are linearly combined to form a single basis for decision." The prediction of performance based on this assumption corresponds to Equation 9 or 11. We saw that this model is consistent with optimal behavior if a set of rather strong assumptions regarding the noise process are valid. Otherwise, the TSD model has no normative justification. Of course, it nevertheless may turn out to be empirically useful for explaining behavior.

Linear Integration Model (LIM)

Anderson (1981, 1982) and his students have established the most comprehensive framework for the analysis of integration. This framework is called *information integration* and uses the tools of functional measurement—most notably, analysis of variance (ANOVA) and interval-response scales. In a seminal study, Anderson (1962) initiated this methodological and theoretical framework for the study of person impression (Asch, 1964). Methodologically, a factorial design was used to indepen-

dently vary descriptive adjectives of a hypothetical person. Anderson used three adjective factors with three levels along each factor, giving a total of 3^3 or 27 unique adjective combinations. A subject was tested repeatedly on each of the 27 unique descriptions presented in a random order. The three levels along each factor contained adjectives of high, medium, and low likableness value. On a given trial, a subject might judge a hypothetical person who was good-natured, unsophisticated, and tactful. The judgments involved a 20-point rating along a scale between *likable* and *dislikable*. An ANOVA was performed on the judgments of individual subjects to assess the contribution of each factor and any interaction among the factors. As expected, there were large effects of likableness value for each factor, but, surprisingly, there was no interaction among the factors for 9 of the 12 subjects. The interaction for the other 3 subjects was relatively small and accounted for very little of the variance.

Implementation of Adding Rule

According to an adding model, evaluation (called *valuation* by Anderson) involves the processes that transform the physical stimulus to its psychological representation (Anderson, 1981). Integration involves a linear combination of scale values made available by evaluation. The decision (called *response function*) is assumed to be linear; that is, the integrated value can be mapped linearly into a rating scale. This decision process is equivalent to the RGR. For categorical responses, either the RGR or a CR is assumed.

We first derive the predictions for the addition of values representing the different sources of information along with the RGR. The integration is computed by the addition of the values representing the evaluation of each source of information (Anderson, 1965; Anderson & Cuneo, 1978). If c represents the degree to which the oval is closed and o represents the degree to which the straight line is oblique, the outcome of integration would be:

$$a_Q = c + o$$

and

$$a_G = (1 - c) + (1 - o).$$

If Q and G are the only valid response alternatives, the decision operation would determine their relative merit under the RGR, leading to the prediction that both the two-choice classification judgments and the rating judgments would be equal to

$$P(A_Q|C = C_i, O = O_j) = \frac{c + o}{c + o + [(1 - c) + (1 - o)]} = \frac{c + o}{2}, \tag{12}$$

where $P(A_Q|C = C_i, O = O_j)$ would be the proportion of Q judgments or a rating of Q-ness on a scale of 0 to 1, given the test letter C_iO_j.

Implementation of Averaging Rule

An averaging rule derived from the domain of personality impression is a viable and intuitively plausible candidate for pattern recognition and decision making (Anderson, 1973). Given continuous and independent evidence from the information sources, the perceiver might simply average the sources of evidence and classify or rate the pattern on the basis of the computed average. Given the averaging rule, the Q-ness of a test letter, a_Q, can be assumed to be an average of its two component features:

$$a_Q = \frac{c + o}{2}. \tag{13}$$

An extension of the averaging rule is a weighted averaging rule, in which one of the features would receive more weight than the other (Anderson, 1981; Massaro, 1985). For example, the oval might contribute more to the judgment than the line. In the present formulation of the model, however, the scale values may be viewed as already incorporating weights so that the two models are not identifiably different. Although the generalized TSD also assumes integration by a weighted averaging process (Anderson, 1974), the c_i and o_j values are first subjected to a $Z(\bullet)$ transform, and their weighted average is subjected to an $N(\bullet)$ transform.

In Anderson's theory of averaging, no explicit decision stage was deemed to be necessary given that the rating judgment was taken to be a direct reflection of outcome of the integration process. At first glance, this assumption seems reasonable when graded rating judgments are used. As noted by Anderson (1974), a discrete judgment would necessarily demand an explicit decision operation. Once the operation is admitted for discrete judgments, it might be argued that it is also involved in continuous rating judgments. What is revealing in this regard is how the explicit decision operation changes the interpretation of the averaging results observed by Anderson and others. Comparing Equations 12 and 13, we see that the results of averaging imply an additive integration rule when the model is implemented with the RGR for the decision stage.

Optimality Properties of LIM and Relation to TSD

The adding rule with an RGR and the averaging rule are nonoptimal models of information integration. The response given two sources of information supporting the same alternative is a compromise between the responses given to the separate sources presented in isolation. Optimal integration (i.e., Bayes's theorem) dictates that the response given two independent sources be more extreme than either of the responses given the separate sources supporting the same alternative. According to optimal integration, our opinion of someone should always become more favorable with additional positive information, even if the new information is not as favorable as some of the old. Averaging, on the other hand, predicts that our overall opinion is diminished if the new positive information is less positive than the old.

In the context of general categorical-response experiments, Anderson (1974) appears to have viewed his algebraic (linear) integration model as conceptually equivalent to the extension of TSD we described in the graded factorial designs section. In particular, his Equation 18, which incorporates a CR decision process, coincides with our Equation 11 in the case of equal weights and an unbiased criterion. Hence, we regard the CR version of the linear integration model (LIM–CR), in its adding, av-

eraging, or weighted versions, as observationally equivalent to our extension of the TSD endowed with the same number of free parameters, for any categorical response experiment. Given an interpretation of responses as subjective probabilities, this algebraic (linear) integration model is inefficient (nonoptimal) except in the special case of evaluation degraded by equal-variance normal noise processes.

Two-Layer Connectionist Model of Perception (CMP)

There has been a tremendous revival of models based on the metaphor of neural information processing. In these connectionist models, information is represented in terms of the activations and inhibitions of neurallike units (Minsky & Papert, 1969/1988; Rosenblatt, 1958; Rumelhart & McClelland, 1986). These units are assumed to exist at different layers; for example, the TRACE model of speech perception (McClelland & Elman, 1986) consists of units at the feature, phoneme, and word levels. The units interact with one another via connections with positive or negative weights that are either specified in advance or learned through feedback.

Numerous layers and adjustable weights make possible many varieties of connectionist models (see Golden, 1988, for a partial survey). We consider here only a specific two-layer connectionist model of perception (CMP) that is most comparable with the alternative models such as the TSD and FLMP. The two layers correspond to an input and an output layer. Connectionist models with more than two layers may be more powerful but are usually much less parsimonious; that is, they require many free parameters. Models with an intermediate (hidden) layer of units, for example, can describe results that are not linearly separable (Massaro, 1988b). In effect, a hidden layer of units violates our independence assumption in information evaluation and falls outside our conceptual framework.

The CMP is assumed to have input and output layers of neural units, with all input units connected to all output units. For ease of exposition, we assume that each level of each source of information is represented by a unique unit at the input layer. Each response alternative is represented by a unique unit at the output layer. Figure 8 gives a schematic representation of two input units connected to two output units.

An input unit has zero input, unless its corresponding level of the stimulus dimension is presented. This constraint ensures that only one input unit is activated per given presentation of a source of information. Presentation of an input unit's target stimulus gives an input of 1. The activation of an output unit by an input unit is given by the multiplicative combination of the input activation and a weight w. With two active inputs X_i and Y_j, the activation entering output unit a_1 is $x_1 + y_1$, where $x_1 = wX_i$ and $y_1 = vY_j$. Analogous to the use of negation in the FLMP, the weight on the activation entering output unit a_2 can be assumed to be the negative of the weight entering a_1 (Massaro & Cohen, 1987). In this case, the activation entering output unit a_2 is $x_2 + y_2$, where $x_2 = -wX_i$ and $y_2 = -vY_j$. The total activation leaving an output unit is given by the sum of the input activations passed through a sigmoid-squashing function (Rumelhart, Hinton, & Williams, 1986). Therefore, for an X_iY_j stimulus,

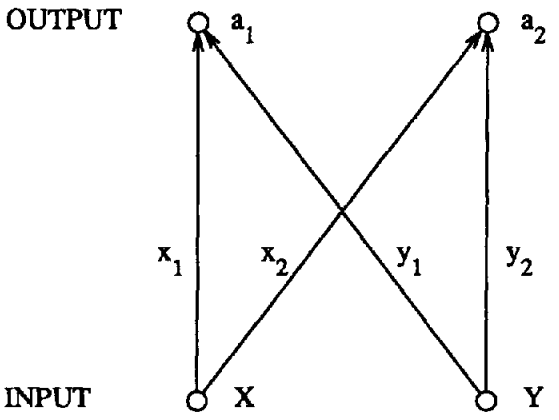

Figure 8. Illustration of connectionist model with two input units, X and Y, and two output units, a_1 and a_2. (The activations entering a_1 from X and Y are x_1 and y_1, and analogously for a_2.)

$$a_1 = \frac{1}{1 + e^{-[x_1+y_1]}} = \frac{1}{1 + e^{-[x+y]}},$$

and

$$a_2 = \frac{1}{1 + e^{-[x_2+y_2]}} = \frac{1}{1 + e^{-[(-x)+(-y)]}}.$$

The neural processing of a connectionist model does not specify completely the stimulus–response function. The activations at the output layer have to be mapped into a response, and an RGR is usually assumed to describe this mapping (McClelland & Elman, 1986). Taking this tack, the activation a_1 transformed into a response probability by the RGR gives

$$P(A_1|XY) = \frac{\dfrac{1}{1 + e^{-[x+y]}}}{\dfrac{1}{1 + e^{-[x+y]}} + \dfrac{1}{1 + e^{-[(-x)+(-y)]}}}. \qquad (14)$$

In summary, evaluation in the CMP consists of the activation of neurallike units. Integration involves the summation of the separate activations passed through the sigmoid-squashing function. Decision follows the RGR.

Implementation of the CMP

Given a test letter in our prototypical task, there are two active input units corresponding to the closedness and obliqueness dimensions. The CMP does not specify the psychophysical relationship between the physical stimulus and its sensory transformation. Analogous to the other models, the CMP requires free parameters to specify this relationship. A unique weight is assumed for each level of each source of information in the CMP. The number of free parameters is equal to the number of levels of the closedness dimension plus the number of levels of the obliqueness dimension. Although an additional threshold unit is sometimes assumed in connectionist models, no such unit is assumed herein.

Presentation of a test letter would activate two input units, corresponding to the appropriate levels of the obliqueness and

closedness dimensions. Presentation of an input unit's target stimulus gives an input of 1; otherwise, the input unit has an input of 0. For a test letter in our prototypical task, the activation of output unit by an input unit is given by the multiplicative combination of the input activation and a weight. The activation entering the output unit corresponding to the alternative Q would be $c + o$, where $c = wC_i$ and $o = vO_j$. Given our negation normalization, the activation entering output unit corresponding to the alternative G would be $-(c + o)$. Thus, the total activation leaving output units Q and G are

$$a_Q = \frac{1}{1 + e^{-[c+o]}}$$

and

$$a_G = \frac{1}{1 + e^{-[(-c)+(-o)]}}.$$

These activations are transformed into response probabilities by the RGR so that

$$P(Q|C_iO_j) = \frac{\dfrac{1}{1 + e^{-[c+o]}}}{\dfrac{1}{1 + e^{-[c+o]}} + \dfrac{1}{1 + e^{-[(-c)+(-o)]}}}. \qquad (15)$$

Comparison of the CMP and FLMP

A comparison between the FLMP and CMP reveals that the two models, couched in different theoretical frameworks, can make identical predictions in practice (Massaro & Cohen, 1987). In this case, a formal equivalence between the two models exists if adding the weighted activations at input and transformed by the sigmoid-squashing function is mathematically equivalent to multiplying fuzzy-truth values. We now demonstrate that such an equivalence holds in the case of a simple categorical-response experiment. That is, if an experiment allows subjects only two response alternatives, then in their standard implementations, the CMP and FLMP are observationally equivalent.

For notational simplicity, we consider two sources of information, denoted X and Y; the argument remains valid but requires more complex notation when more than two information sources are present. Let P_i (or Q_j) denote the observed response probabilities for the first response alternative (e.g., the G response in our prototypical task) in single-factor trials with X set at level i (or Y set at level j). Let $S(t) = 1/(1 + e^{-t})$ denote the sigmoid-squashing function mapping $t \in (-\infty, \infty)$ to $u \in (0, 1)$, and let $S^{-1}(u) = -ln(1/u - 1)$ denote its inverse. Note that the weights $V_i = S^{-1}(P_i)$ and $W_j = S^{-1}(Q_j)$ will be chosen for the CMP from data that generate P_i and Q_j under the convention that input units have a value of 1 if activated and 0 if not activated. In view of Equations 10 and 14, the demonstration reduces to verifying that

$$\frac{P_iQ_j}{P_iQ_j + (1 - P_i)(1 - Q_j)} = \frac{S(V_i + W_j)}{S(V_i + W_j) + S(-V_i - W_j)} \qquad (16)$$

for all values of P_i, Q_j in [0, 1]. First note that $S(t) + S(-t) = 1$ for all t because multiplying both the numerator and denominator by e^t gives

$$S(t) + S(-t) = \frac{1}{1 + e^{-t}} + \frac{1}{1 + e^{+t}} = \frac{e^t}{e^t + 1} + \frac{1}{e^t + 1} = 1.$$

It follows that the denominator of the right-hand side of Equation 16 is equal to 1, so we can expand the right-hand side as follows:

$$S(V_i + W_j) = \frac{1}{1 + e^{-V_i - W_j}} = \frac{1}{1 + e^{-V_i}\left(\dfrac{1 + e^{V_i}}{1 + e^{V_i}}\right)e^{-W_j}\left(\dfrac{1 + e^{W_j}}{1 + e^{W_j}}\right)}$$

$$= \frac{1}{1 + \left(\dfrac{1 + e^{-V_i}}{1 + e^{V_i}}\right)\left(\dfrac{1 + e^{-W_j}}{1 + e^{W_j}}\right)}$$

$$= \frac{\left(\dfrac{1}{1 + e^{-V_i}}\right)\left(\dfrac{1}{1 + e^{-W_j}}\right)}{\left(\dfrac{1}{1 + e^{-V_i}}\right)\left(\dfrac{1}{1 + e^{-W_j}}\right) + \left(\dfrac{1}{1 + e^{V_i}}\right)\left(\dfrac{1}{1 + e^{W_j}}\right)}$$

$$= \frac{S(V_i)S(W_j)}{S(V_i)S(W_j) + [1 - S(V_i)][1 - S(W_j)]},$$

which of course corresponds to the left-hand side of Equation 16, and the verification is complete.

It is important to recognize that the observational equivalence between CMP and FLMP as models of information integration does not extend to experiments allowing more than two response alternatives. For example, suppose that the probabilities of responses A_1, A_2, and A_3 are .6, .2, and .2, respectively, given information source X (at some specified level i) in isolation, whereas the corresponding probabilities for Y are .7, .1, and .2. Then the weights for X are $S^{-1}(.6) \approx .405$, $S^{-1}(.2) \approx -1.386$, and $S^{-1}(.2) \approx -1.386$, whereas the weights for Y are .847, −2.197, and −1.386. The CMP prediction for response A_1 given both sources is

$$\frac{S(.405 + .847)}{S(.405 + .847) + S(-1.386 - 2.197) + S(-1.386 - 1.386)}$$

$$= \frac{.778}{(.778 + .027 + .059)} = .9004.$$

The FLMP prediction is

$$\frac{(.6)(.7)}{(.6)(.7) + (.2)(.1) + (.2)(.2)} = .8750.$$

Although the difference between these predictions is not striking in this example, it does establish the nonequivalence of the two models for three or more response alternatives.

Multidimensional Scaling (MDS)

A related but different attack on the problem of assessing the influence of multiple sources of information is multidimensional scaling (MDS), developed by Shepard (1962, 1988), his colleagues, and others (Kruskal, 1964). MDS has been applied to both similarity judgments and recognition judgments. Tradi-

tionally, these researchers have not manipulated the properties of test objects in pattern-recognition tasks. Usually, investigators use only the endpoint categories (a categorical design in our framework) and examine the pattern of errors that subjects make when they identify the patterns (Bouma, 1971; Loomis, 1982; Shepard, 1988). To induce a reader to make errors, for example, letter stimuli are degraded by presenting them for a short duration or at a great distance. The responses of the subjects are entered into a confusion matrix that indicates the identification frequencies for each letter stimulus. For example, subjects might be given the set of 26 lowercase letters in the English alphabet and respond with one of the 26 alternatives on each trial. These results are then used to distinguish among various descriptions of the properties of the letters. The goal has been to find the smallest number of dimensions that best describes the responses (Gilmore, Hersh, Caramazza, & Griffin, 1979; Shepard, 1988; Townsend, 1971). The usual MDS approach differs from the other approaches to information integration, which all specify the sources of information in advance, and then describe various ways in which the values obtained from the evaluation of each source are combined. In its standard application, MDS envisages the reverse process: One seeks to infer the number of independent sources (and perhaps their specification) from an analysis of response data.

Nevertheless, MDS can also be implemented as a model of information integration, and indeed, Ashby and Gott (1988, p. 34) do so explicitly (see also Ashby & Perrin, 1988). To implement MDS, one takes the identity (and number, that is, dimensionality) of the sources of information as given, and regards the evaluation of all information sources for some stimulus as defining a point in a vector space of the given dimension. Each possible response is defined as a point in the same vector space. Given a distance function for the vector space such as the Euclidean or the "city-block" metric, one assumes a decision rule based on minimum distance between the point and the prototype: The individual chooses the response nearest the evaluated stimulus. The result is an integration model that uses a distance metric as its integration function. Such models are particularly well adapted to discrete-response experiments using a factorial design. We take the liberty in what follows of referring to them as *MDS models*.

Implementation of MDS

Assume that n independent variables, or sources of information, are used, and construct a vector space of dimension n with axes that refer to evaluation scale values for these sources. Suppose also that the design allows several responses, each of which can be assigned to a point in this vector space. For example, in our G–Q recognition task, we have a two-dimensional space in which the horizontal axis measures the degree of closedness of a circle and in which the vertical axis measures the degree of obliqueness of the line. If the allowable responses consist only of G and Q, then the two allowable responses might have recognizable locations at the points $A_G = (10, 0)$ and $A_Q = (0, 61)$ for G and Q, respectively. In the simplest implementation of MDS, we take the evaluation process to be essentially a noiseless scaling of the two sources of information, so x = degree of closedness and y = obliqueness of line in degrees from horizontal of the stimulus presented. Alternatively, one can assume that some specified noise process degrades evaluation; for example, multivariate normal noise in the general Gaussian model of Ashby and Perrin (1988).

To define an MDS integration function, we need to specify a metric, or distance function, on the vector space. All examples in the MDS literature use the Minkowski r metric, defined for pairs of n vectors $x = (x_1, \ldots, x_n)$ and $A = (A_1, \ldots, A_n)$ by the formula

$$|x - A|_r = [\sum_{i=1}^{n} |x_i - A_i|^r]^{1/r}, \tag{17}$$

where the exponent r is a number between 1 and ∞. Observe that in the $r = 1$ case, the vector distance between two points is the sum of the component factor distances. This case is known as the *city-block metric* because the overall distance one must cover when traveling on a grid of city streets is the sum of the north–south and east–west distances. The case most often encountered, $r = 2$, is known as Euclidean distance, because (according to Pythagoras's ancient theorem) it measures distance "as the crow flies" in standard (Euclidean) geometry. Sometimes positive weights w_i are assumed to multiply the terms in Equation 17, but little further generality is so achieved: The same result can be achieved by changing the scale i in proportion $w_i^{1/r}$, that is, by a change in units for each information source. For specified r, the integration function in MDS is given by $a_k = |x - A_k|_r$, where A_k is the vector corresponding to response alternative k, and x is the vector defined by the evaluated stimulus.

One possible decision rule in an MDS model is a generalization of the CR: response k is selected if $a_k = \min\{a_1, \ldots, a_n\}$. That is, we assume that an individual selects the response alternative closest to the perceived stimulus. Recall that the basic CR in a simple, unbiased two-response case defines a point that is equidistant from the two alternatives. In the present case of n dimensions, this would correspond to the locus of points equidistant from the two response alternatives. For example, in the G–Q letter-recognition task, the $r = 2$ (Euclidean distance) metric defines the perpendicular bisector of the line segment connecting the points $A_G = (10, 0)$ and $A_Q = (0, 61)$ as the generalized criterion: Evaluated stimuli that fall on the Q side of this bisector generate Q responses, and evaluated stimuli on the other side generate G responses.

The separating boundaries between the response regions can be more complex than straight lines (or $n - 1$ dimensional hyperplanes for n information sources). For example, in Ashby and Gott's (1988) general Gaussian model, the boundaries are conic sections. Even the simplest city-block case has boundaries that consist of three connected lines, two oriented along an axis and the third a connecting diagonal. However, in every case, the generalized CR partitions the vector space into m regions, one for each allowable response, and evaluated stimuli that lie in a given region all produce the same response.

To summarize, MDS can be implemented as an information-integration model in which the currency is distance. The evaluation of stimuli can be assumed to be noiseless or to be degraded by a specific noise process. The integration function is defined by the Minkowski r metric for some specified r. The decision rule is usually a generalized CR.

Table 1
Summary of the Currency and Processes Assumed by the Integration Models

Model	Currency	Evaluation	Integration	Decision
FLMP	Truth values	Noise free	Multiplication	RGR
TSD	Sensory information	Normal noise	Summation	CR
LIM-RGR	Valuations	Noise free	Addition	RGR
LIM-CR	Valuations	Normal noise	Addition	CR
CMP	Activations	Noise free	Addition/sigmoid transform	RGR
MDS	Distance	Normal noise	Euclidean or city-block metric	CR

Note. FLMP = fuzzy-logical model of perception; TSD = theory of signal detectability; LIM-RGR = linear integration model–relative goodness rule; LIM-CR = linear integration model–criterion rule; CMP = connectionist model of perception; MDS = multidimensional scaling.

Relation to Other Integration Models

Clearly, one implementation of MDS is closely related to the TSD model. Indeed, one obtains precisely the standard TSD model that we presented earlier under the following assumptions (Ashby & Gott, 1988): (a) Noise at evaluation comes from the same multivariate normal distribution for each stimulus, the distribution having a mean and correlation of 0 across factors; (b) the integration function is Euclidean ($r = 2$) distance; and (c) the decision rule is generalized criterion (the closest response alternative is always chosen). This equivalence of MDS and TSD provides a geometric interpretation of the key Equation 9 for TSD: The d's for each factor represent distances along perpendicular axes, and the overall distance for two factors is the length of the hypotenuse, so Equation 9 is just the Pythagorean theorem.

Another implementation of MDS is based on a city-block integration function ($r = 1$). This model turns out to be equivalent to a TSD model in which the separate d's are added. As Fidell (1970) pointed out, noise processes that are perfectly correlated across factors (or sources) leads to an overall d' that is the sum of the individual factor d's, as in the city-block ($r = 1$) metric. If the d's arise from logistic rather than normal noise, then the city-block MDS model appears to be equivalent to the FLMP for two response alternatives.

In conclusion, the MDS approach yields some valuable insights into the geometry of information integration and allows several integration models to be constructed once the r metric (and the noise process at evaluation and the decision rule) are specified. However, the two most natural specifications yield a model equivalent to TSD and one similar to the FLMP. The more general specifications introduced by Ashby and Perrin (1988) involve many additional free parameters (e.g., for the covariance matrix). For present purposes, then, MDS does not provide any additional simple models to be compared with those already introduced, and thus MDS models are not included in our empirical assessment of models of integration.

Empirical Predictions and Tests of the Models

In the last five sections, we have developed several models of information integration. The critical features of the models are summarized in Table 1. Our analysis revealed differences and similarities among the models. Given two response alternatives, both the CR implementation of the LIM (LIM–CR) and the basic Euclidean version of MDS are observationally equivalent to the TSD. Although the FLMP and CMP are observationally equivalent for two response alternatives, they differ for three or more response alternatives. Hence, four distinct models remain for comparison: FLMP, TSD, LIM–RGR, and CMP. We have already discussed the optimality properties and demonstrated the mathematical nonequivalence of these models. Given this set of plausible models, the real basis of comparison is the predictive power of the models. Reliable assessments will be possible only after the models have been contrasted in a broad range of experimental tasks. We initiate this project by generating specific predictions and providing some simple illustrative empirical comparisons.

Hypothetical predictions were generated from each model for the results of an expanded two-factor design with two and with four categorical-response alternatives. Recall that the expanded design tests each of the two sources of information presented in isolation, as well as the factorial combination of the two sources of information. The design provides a more powerful data base to assess models of human performance than do standard factorial designs (Massaro, 1987b). There were seven levels of each of the two independent variables. To generate each model's predictions, hypothetical parameter values were assigned to each of the single-source conditions. These values are given in Table 2. The hypothetical parameter values in Table 2 were chosen to be asymmetric around .5 and to cover different ranges between 0 and 1. (Curves generated from symmetric parameter values are redundant, and real stimulus continua seldom turn out to be symmetric or to cover the same range.) Each model predicts that the probability of a particular response is some combina-

Table 2
Hypothetical Parameter Values for the 14 Single-Source Conditions for Models with Expanded Two-Factor (X and Y) Design

	Level						
Factor	1	2	3	4	5	6	7
X	.01	.10	.30	.50	.70	.90	.99
Y	.03	.20	.40	.60	.80	.92	.95

tion of unique parameter values associated with each of the levels of the two independent variables.

The predictions of the models under consideration can be fit to data with a parameter-estimation program such as STEPIT (Chandler, 1969). A model is defined in STEPIT as prediction equations that contain a set of unknown parameters. STEPIT minimizes the deviations between the observed and predicted values of the models by iteratively adjusting the parameters of the equations. Root mean square deviation (RMSD) values index the overall goodness of fit of the model, and their use ensures a maximum likelihood fit. The RMSD value is the square root of the average squared deviation between the predicted and observed values. The smaller the RMSD value, the better the fit of the model.

RMSD values are used because these specify directly the correspondence between a model and data or the correspondence between the predictions of two models. That is, an RMSD value of 0.05 means that the observations and predictions are within roughly 0.05 of one another on the average. More important, we are evaluating similarities and differences among different models from which predictions are in terms of probabilities, not actual frequencies. Other measures of goodness of fit, including chi-square, require knowledge about the actual frequencies in each cell. We know that with large enough frequency, any model—no matter how good the fit to data—can be rejected. Although other statistical tests might be useful in other contexts, an RMSD goodness of fit seems most appropriate for our purposes.

Two Response Alternatives

We first consider the case in which there are two possible responses in the task. We generated hypothetical data as follows. For all five of the models, the probability of an A_k response, $P(A_k)$, to the single-source conditions was assumed to be equal to the corresponding parameter value in Table 2. The $P(A_k)$ values to the factorial conditions were then generated from the values in Table 2. The currency of the FLMP and LIM–RGR are values between 0 and 1, and their predictions follow directly from these values. For the TSD and LIM–CR, the parameter values must be transformed into z scores before integration, and the outcome of integration must be transformed back into $P(A_k)$ values. The currency of the CMP is activation weights that can vary between large negative and large positive values, but after the sigmoid transformation, we again obtain normalized values between 0 and 1. Given the constraints on the generation of the hypothetical results, identical predictions are made by all of the models for the single-source conditions. The similarities and differences among the models can thus be seen directly by contrasting the predictions for the factorial condition.

Evaluating how much the five models differ from one another is informative. Logically, one model might mimic the results of another simply with a change in parameter values. To explore this issue, the five models were fit to the five sets of predictions generated by these same models (Table 3). In all cases, 14 parameter values (2 variables × 7 levels for each variable) were estimated to minimize the RMSDs between the observed and predicted data.

Each model can describe data generated by itself and by models that are mathematically equivalent to it. As expected from the mathematical analyses of the models, the FLMP and CMP made identical predictions to one another, as did TSD and LIM–CR (Table 3). The LIM–RGR predictions were unique. Thus, there are three different sets of predictions. The predictions for the factorial condition by these three classes of models are given in Figure 9. As can be seen in the figure, the three classes of models make noticeably different predictions from one another. Especially noticeable is the difference between the LIM–RGR and the other two classes of models. Linear integration followed by the RGR produces additive results that plot as parallel curves. The predictions for the other two classes of models are clearly elliptical, with the distances among the curves much greater in the middle of the range of parameter values than at the extremes. Even the other two classes differ significantly, however, in the fine structure of their predictions. The FLMP and CMP class is more continuously graded across the continuum relative to the TSD and LIM–CR class. The three classes of models shown in Figure 9 are identifiably different from one another. That is, the models cannot accurately describe predictions generated by each other by simply assuming another set of parameter values. It is not possible to find a set of parameter values for one model that will produce predictions that will mimic the results generated by another model. For the identifiably different models, the RMSD values are sufficiently large to warrant the belief that these models could be distinguished from one another in practice.

Table 3
Root Mean Square Deviation Values for Fits of the Five Models to the Five Sets of Predictions

	Model				
Data	LIM–RGR	FLMP	CMP	TSD	LIM–CR
LIM–RGR	.000	.068	.068	.084	.084
FLMP	.159	.000	.000	.041	.041
CMP	.159	.000	.000	.041	.041
TSD	.191	.033	.033	.000	.000
LIM–CR	.191	.033	.033	.000	.000

Note. The predictions are for an expanded two-factor design with two response alternatives given the parameter values in Table 2. LIM–RGR = linear integration model–relative goodness rule; FLMP = fuzzy-logical model of perception; CMP = connectionist model of perception; TSD = theory of signal detectability; LIM–CR = linear integration model–criterion rule.

The models were also tested against real data from the Massaro and Hary (1986) task described in the Taxonomy of Experiments section. Table 4 gives the RMSD values. As expected, the FLMP and CMP gave equivalent descriptions, as did the TSD and LIM–CR models. Figure 10 gives the observed results along with the predictions of the three classes of models. As can be seen in the figure, the LIM–RGR gives a poor description of the results relative to the good description of the FLMP and CMP and the TSD and LIM–CR classes of models. Although the fit of the latter two classes of models were both fairly good, an ANOVA performed on the RMSD values revealed that the FLMP and CMP class of models gave a significantly better fit of the results than did the TSD and LIM–CR class of models, $F(1, 8) = 79.46$, $p < .001$.

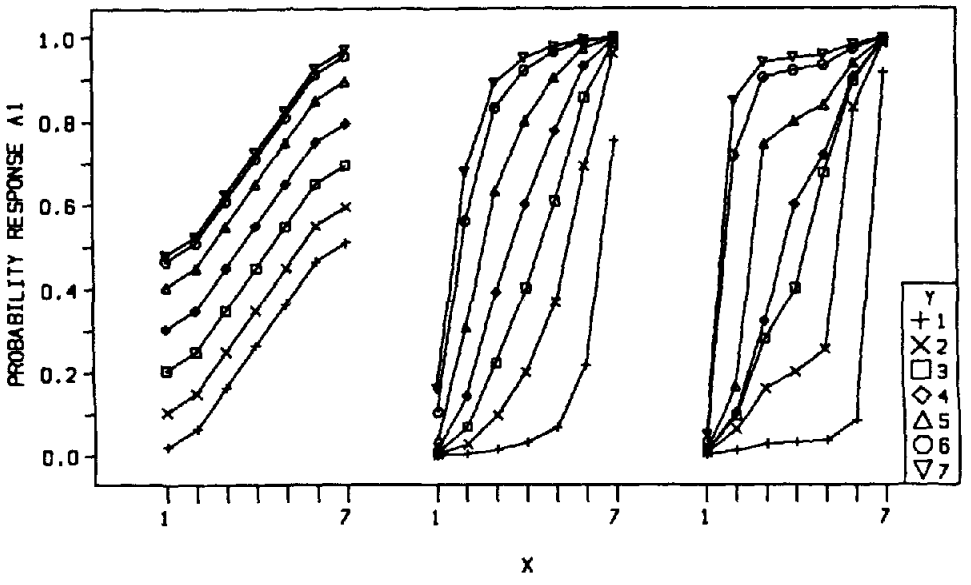

Figure 9. Predicted probability of *A*1 responses for the factorial conditions of the expanded factorial design given the parameter values in Table 1. (The left panel gives the predictions for the linear integration model–relative goodness rule, the center panel gives the predictions for the fuzzy-logical model of perception and connectionist model of perception, and the right panel gives the predictions for the theory of signal detectability and linear integration model–criterion rule.)

Graded Responses

The predictions of the models for graded responses are identical to those predicted for categorical responses. Thus, the predictions shown in Figure 9 can be tested against both categorical and graded responses (assuming that subjects use a linear response scale in the graded task). In Massaro and Hary's (1986) rating task, 6 subjects rated the *Q*-ness–*G*-ness of a test letter from the *G*–*Q* continuum by using a rating scale displayed on the computer terminal monitor. The scale was a straight horizontal line made up of 51 divisions, although it was displayed as a continuous line on the monitor screen. The left end of the scale was labeled *Q* and the right end *G*. Subjects were able to move a pointer along the scale but were not told that the scale had 51 divisions. The pointer was represented as a black box on the rating scale, and subjects manipulated the pointer using left and right arrow keys on the terminal keyboard. Subjects were instructed to "tell us where the test letter falls on the scale from *Q* to *G* by moving the pointer on the screen in front of you. . . . We want you to use the whole *Q*–*G* scale to respond with, not just the two endpoints and middle, for example. For the letters you will see in this study, you should use the entire scale and all of the points in it."

Table 4
Root Mean Square Deviation Values for the Fits of the Three Classes of Models to the G–Q *Categorical-Response Task of Massaro and Hary (1986)*

	Model		
Subject	LIM-RGR	FLMP/CMP	TSD/LIM-CR
1	.247	.035	.032
2	.154	.064	.072
3	.157	.097	.110
4	.202	.044	.048
5	.226	.031	.044
6	.229	.054	.050
7	.236	.049	.057
8	.238	.028	.030
9	.192	.054	.081
M	.209	.051	.058

Note. LIM-RGR = linear integration model–relative goodness rule; FLMP/CMP = fuzzy-logical model of perception and connectionist model of perception; TSD/LIM-CR = theory of signal detectability and linear integration model–criterion rule.

With the assumption of a linear response scale, the rating task provides a direct test between linear and nonlinear integration. Linear integration (Anderson, 1981, 1982) makes strong predictions about the average rating response in an integration task: If a subject rates a test letter on an interval scale that varies on two factors, then the plot of the ratings versus the factors should produce parallel lines. The additive rule assumes that the contribution of one factor to integration is the same regardless of the ambiguity of the other factor. This rule is not optimal in that averaging an ambiguous source of information with an informative source will tend to neutralize the judgment relative to the informative source presented alone. In contrast, the FLMP predicts American-football-shaped curves when the average ratings are plotted in a two-factor graph. These curves reflect the larger impact of the less ambiguous source of information. A test between these different predictions was carried out by fitting the respective models to the individual rating judgments

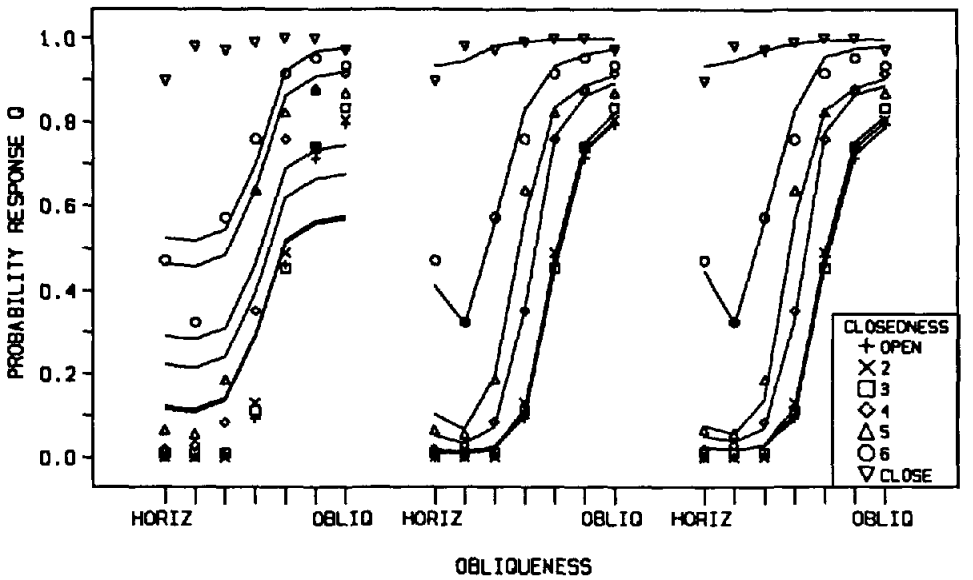

Figure 10. Observed (points) and predicted probability of a *Q* identification as a function of the closedness of the gap and the obliqueness of the line (after Massaro & Hary, 1986). (The left panel gives the predictions for the linear integration model–relative goodness rule, the center panel gives the predictions for the fuzzy-logical model of perception and connectionist model of perception, and the right panel gives the predictions for the theory of signal detectability and the linear integration model–criterion rule.)

of the Massaro and Hary (1986) study. Figure 11 illustrates the model fits averaged over subjects. The parallel lines predicted by the LIM–RGR do a rather poor job in fitting the data points. The FLMP does much better than the additive model. The RMSDs for the individual subjects are presented in Table 5.

Four Response Alternatives

Given that some of the models make mathematically equivalent predictions in tasks with two response alternatives, tasks with a larger number of alternatives need to be considered to

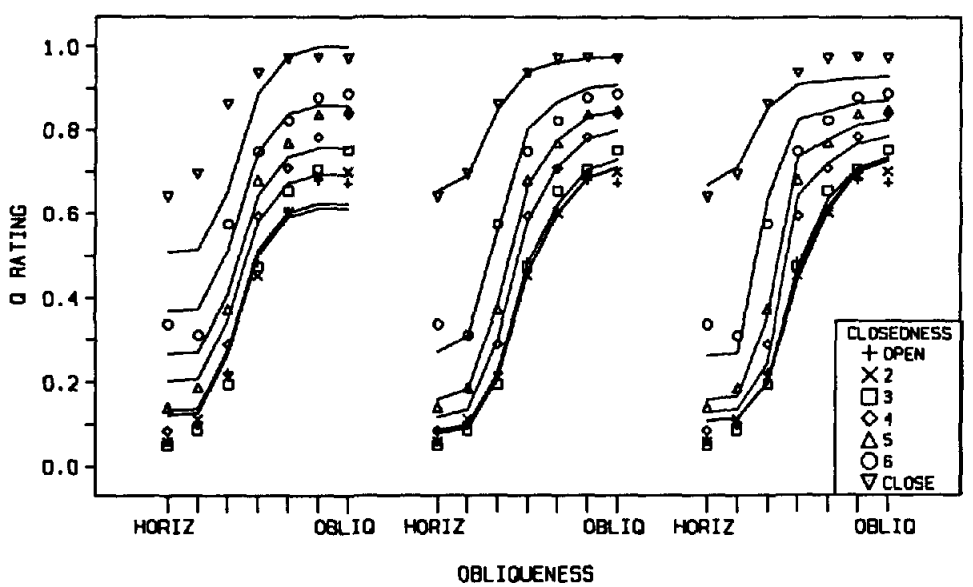

Figure 11. Observed (points) and predicted *Q* rating as a function of the closedness of the gap and the obliqueness of the line (after Massaro & Hary, 1986). (The left panel gives the predictions for the linear integration model–relative goodness rule, the center panel gives the predictions for the fuzzy-logical model of perception and connectionist model of perception, and the right panel gives the predictions for the theory of signal detectability and linear integration model–criterion rule.)

Table 5
Root Mean Square Deviation Values for the Fits of the Three Classes of Models to the G–Q *Graded-Response Task of Massaro and Hary (1986)*

	Model		
Subject	LIM-RGR	FLMP/CMP	TSD/LIM-CR
1	.071	.036	.049
2	.164	.037	.047
3	.065	.051	.049
4	.104	.032	.049
5	.083	.042	.057
6	.050	.034	.053
M	.090	.039	.051

Note. LIM-RGR = linear integration model–relative goodness rule; FLMP/CMP fuzzy-logical model of perception and connectionist model of perception; TSD/LIM-CR = theory of signal detectability and linear integration model–criterion rule.

differentiate among these models. To this end, results predicted by the models were also generated for the same expanded two-factor design, but now with four response alternatives. To illustrate this design, we modify the prototypical *G–Q* design in Figure 1 to include the response alternatives *C* and *O*. In this case, the two sources of information are seven levels of closedness of the oval and seven levels of the length of a somewhat oblique line (for example, the third or fourth level of obliqueness illustrated in Figure 1). Given two sources of information, a natural summary description of the four alternatives is

C: not closed oval and no line,

O: closed oval and no line,

G: not closed oval and line,

and

Q: closed oval and line.

If *c* represents the degree to which the oval is closed and *l* the degree to which a straight line is present, the goodness of match with a *C*, *O*, *G*, or *Q* alternative can be represented by the conjunction of these feature values:

$$a_C = (1 - c) \wedge (1 - l)$$

$$a_O = c \wedge (1 - l)$$

$$a_G = (1 - c) \wedge l$$

$$a_Q = c \wedge l,$$

where a_C, a_O, a_G, and a_Q represent the goodness of match of a test letter to the *C*, *O*, *G*, and *Q* alternatives, respectively.

Integration of the two sources of information would give an absolute goodness of match for each of the four alternatives. Decision might consist of either a choice based on a generalized CR or one based on the RGR. With more than two response alternatives, the natural implementation of CR is to choose the alternative with the largest goodness of match. On the other hand, the RGR decision operation determines the relative merit of the alternatives. In the case of RGR, the probability of a response A_C would be equal to

$$P(A_{C|C_iL_j}) = \frac{a_C}{a_C + a_O + a_G + a_Q}, \tag{18}$$

where $P(A_c | C_iL_j)$ is the predicted probability of an A_c response to a particular combination of the two sources of information C_i and L_j.

With seven levels of each factor, an expanded two-factor design with four response alternatives generates 252 data points. These data points were generated with the same parameter values as for two response alternatives (see Table 2). With four response alternatives, the probability of a response given just one source of information was equal to one half the parameter value for that source of information.

Given the constraints on the generation of the hypothetical results, identical predictions are made by all of the models for the single-source conditions. The similarities and differences among the models can thus be seen directly by contrasting the factorial conditions. The FLMP, CMP, and LIM–RGR make straightforward and unique predictions for the four-alternative task. (The implementation of TSD for four alternatives is relatively complex and is not presented herein.) The form of the predictions is apparent in the functions for just one of the four response alternatives. Thus, the predictions of the three models for just one response are given in Figure 12.

As can be seen in the figure, these three models make different predictions from one another. The FLMP predicts a fan-shaped set of curves varying between 0 and 1. The CMP and the LIM–RGR, on the other hand, predict results between 0 and .5. The CMP predicts nonadditive results, whereas linear integration followed by the RGR produces additive results that plot as parallel curves.

The application of these models to a task with four alternatives reveals an important difference between linear and nonlinear integration that was not apparent in the task with just two alternatives. The probability of any response cannot exceed .5 for either the LIM–RGR or the CMP, both of which specify additive integration. Multiplicative integration in the FLMP predicts response probabilities between 0 and 1. The problem with additive integration can be understood by referring to a test stimulus in the hypothetical *QGOC* task. Assume that a source of information gives one unit of support when it matches the alternative and 0 when it does not. If the stimulus is a *C*, then the response alternative *C* receives two units of support. However, the response alternatives *O* and *G* receive one unit of support each for no line and not closed, respectively. That is, with an additive integration rule, each of the *O* and *G* alternatives receives substantial support (approximately half of the support for the alternative *C*). Because the RGR is used for decision, then the probability of a *C* response cannot be greater than the sum of the *O* and *G* response probabilities. In our example, the probability of a *C* response is approximately .5. (This limitation is also true of the TSD model for four responses.) On the other hand, the alternatives *O* and *G* receive little support given multiplicative integration because the poor match on one feature cancels the good match on the other.

As in the case with two alternatives, each model was fit to the predictions of all of the models to address the issue of identifi-

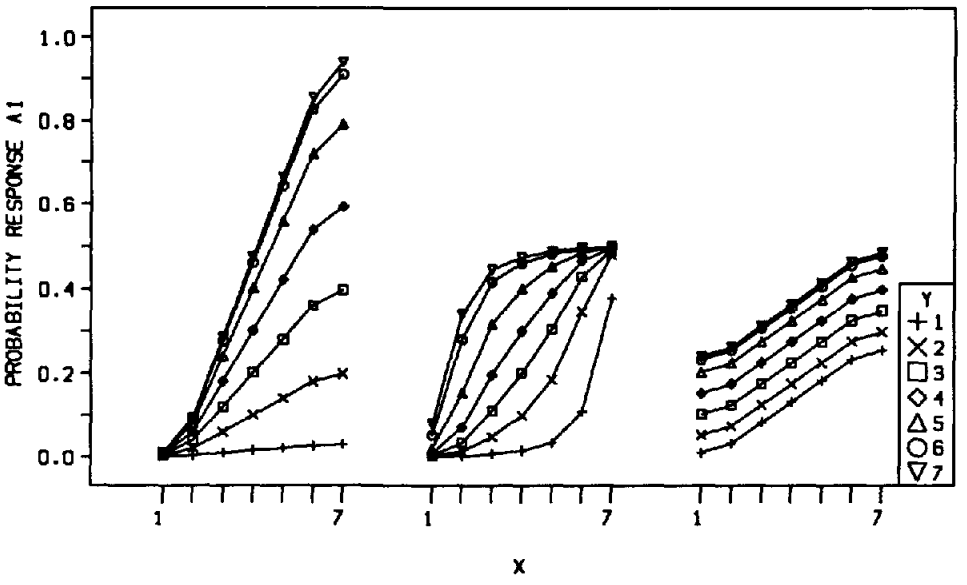

Figure 12. Predicted probability of *A*1 responses for the factorial conditions of the expanded factorial design with four responses given the parameter values in Table 2. (The left panel gives the predictions for the fuzzy-logical model of perception, the center panel give the predictions for the connectionist model of perception, and the right panel gives the predictions for the linear integration model–relative goodness rule.)

ability. Table 6 gives the RMSD values for the fits of the three models to the three sets of data. Each model can describe data generated by itself. The LIM–RGR and CMP are much more similar in their predictions than are the predictions of either of these models to those of the FLMP. The RMSD values are sufficiently large enough to warrant the belief that the FLMP could be distinguished from the LIM–RGR and CMP in practice.

A graded factorial experiment with four response alternatives was carried out by Massaro, Tseng, and Cohen (1983). The four responses in the experiment consisted of four words in Mandarin Chinese. The experimental task was a graded factorial design with seven levels of each of two factors. The factors were the formant structure of the vowel in the monosyllabic words and the fundamental frequency (F_0) contour (tone) during the vowel. Mandarin Chinese is a tone language, and both of these sources of information are functional to distinguish different words. The formant structure was varied to make a continuum of vowel sounds between /i/ and /y/. (The phoneme /y/ is articulated in the same manner as /i/, except with the lips rounded.) The F_0 contour varied from falling–rising to falling during the vowel. Six native Chinese speakers participated for 4 days, giving a total number of 48 responses to each of the 49 test stimuli. The subjects identified each of the 49 test stimuli as one of the four words.

Table 6
Root Mean Square Deviation Values for the Fits of Three Models to the Three Sets of Data Generated with the Parameter Values in Table 2

	Model		
Data	LIM-RGR	FLMP	CMP
LIM-RGR	.000	.024	.011
FLMP	.142	.000	.130
CMP	.087	.082	.000

Note. The data are for an expanded two-factor design with four response alternatives. LIM-RGR = linear integration model–relative goodness rule; FLMP = fuzzy-logical model of perception; CMP = connectionist model of perception.

Figure 13 gives the observed results and the predictions of the FLMP, CMP, and LIM–RGR. Table 7 gives the corresponding RMSD values. As can be seen in the figure, the CMP and LIM–RGR fail catastrophically primarily because they cannot predict a probability of a response greater than .5. The FLMP, on the other hand, captures the results reasonably well. The success of the FLMP is due to the multiplicative integration of the two sources of information. A perfect match of a stimulus with a given response alternative on just one source does not necessarily mean that this alternative should qualify as a reasonably good alternative. Linear integration, however, guarantees that a perfect match of a response alternative with just one source of information will carry significant influence even if the other source of information mismatches the response alternative completely.

In conclusion, we have been relatively successful in testing among the predictions of the models in graded factorial designs with two and four response alternatives and with a graded response. The TSD and LIM–CR class of models and the FLMP and CMP class of models could be discriminated in a graded factorial with just two response alternatives. The LIM–RGR could be rejected in both categorical-response and graded-response tasks. Finally, the FLMP and CMP could be distinguished from

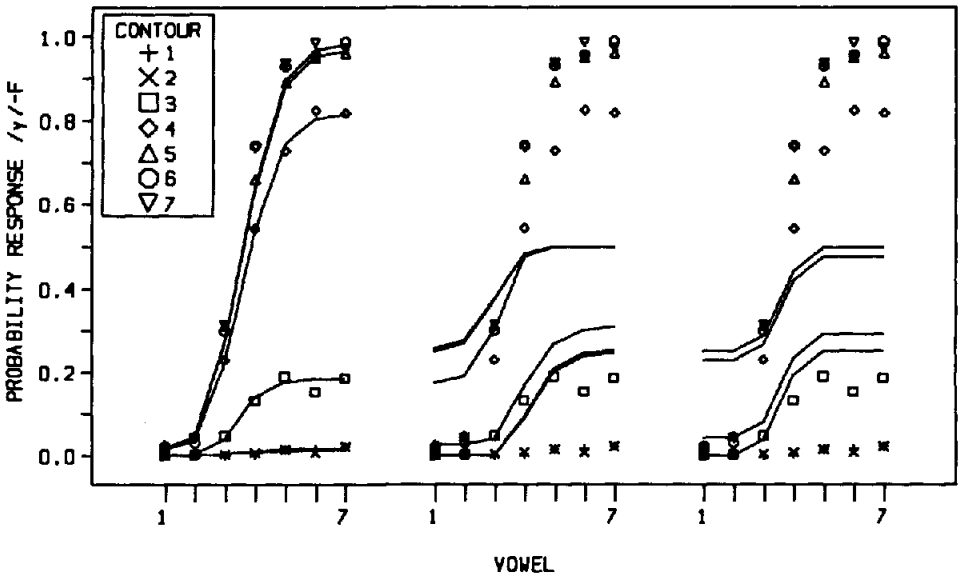

Figure 13. Observed (points) and predicted (lines) probability of /y/-falling responses for the Chinese word identification study (after Massaro, Tseng, & Cohen, 1983). (The left panel gives the predictions for the fuzzy-logical model of perception, the center panel gives the predictions for the connectionist model of perception, and the right panel gives the predictions for the linear integration model–relative goodness rule.)

one another in a graded factorial design with four response alternatives. The FLMP gave a much better description of the results than did the CMP. We caution that the observed advantage of the FLMP is only tentative in that both new results and more refined models could alter the predictive power of the models.

Relation to Other Models

Our presentation of information-integration models is by no means exhaustive. Although we have presented five important models and discussed their optimality and validity properties, we have not covered all variations of the models or discussed other possible models. In the next two sections, we fill in some of these gaps. We then summarize our results.

Table 7
Root Mean Square Deviation Values for the Fits of the Three Models to the Chinese Word-Identification Task of Massaro, Tseng, and Cohen (1983)

	Model		
Subject	LIM-RGR	FLMP	CMP
1	.244	.033	.239
2	.209	.040	.199
3	.227	.030	.220
4	.264	.045	.261
5	.259	.036	.256
6	.248	.072	.242
M	.242	.043	.236

Note. LIM-RGR = linear integration model–relative goodness rule; FLMP = fuzzy-logical model of perception; CMP = connectionist model of perception.

First-Order Versus Second-Order Integration

Shaw (1982) distinguished between first-order and second-order integration models (see also Green & Swets, 1966). In first-order models, the information from the separate sources is integrated prior to making a decision. In second-order models, a categorical decision is made for each source before integration takes place. The separate decisions are then integrated to make a response. The separate decisions in second-order models are categorical and do not preserve the goodness of match of the information leading to the decision. Shaw tested the predictions of these two classes against the results of several different experiments. In the task, one or more stimuli are targets, and a different set of one or more stimuli are nontargets. The stimuli were either brief flashes of light presented to different spatial locations or bursts of sound, and the task was energy detection. In other experiments, the task was letter detection in which the target could appear or not appear in one or more locations. The probability of a detection response under the various experimental conditions was used to test the models. Shaw concluded that second-order decision models gave superior accounts of the results.

The conclusion reached from Shaw's (1982) task appears to conflict with our framework, in which we assumed that all sources of information are integrated prior to any decision. However, an analysis of the experimental tasks reveals that different results in the two domains should not be unexpected. Our tasks involve multiple sources of information specifying the same object. Shaw's tasks, on the other hand, involved deci-

sions about the simultaneous occurrence of multiple objects. That is, targets or nontargets occurring at different locations in the visual display were considered to be different sources of information. Given Shaw's findings, integration of continuous information across objects does not appear to occur in the same manner as integration of information sources specifying the same object. Given multiple objects, subjects apparently categorize each object and then use these categorizations to make a more global decision about the experimental question. That is, an experimental task is not necessarily isomorphic to categorization of an object but could require information derived from multiple categorizations. The *G–Q* task, on the other hand, equates the experimental task with categorization of an object. When this is the case, we expect to find that independent decisions do not occur before integration. Ashby and Gott (1988) also found evidence against independent decisions in a perceptual identification of a horizontal and a vertical line segment attached at an upper left-hand corner. Thus, our conclusions about integration appear to apply to the situation in which multiple sources of information specify a single object. More generally, whether or not integration occurs might be used to determine whether a perceiver treats multiple sources of information as specifying a single object or as specifying multiple objects (Massaro & Cohen, 1988).

Exemplar Models

The models we have discussed in this article belong to a general class of summary-description models, as opposed to exemplar models. Summary-description models are characterized by having each response category defined in terms of a simple conjunction of attributes, features, properties, or dimensions. Exemplar models, on the other hand, define categories in terms of descriptions of several exemplars of the relevant category. The goodness of match of a test item with a category is some function of the goodness of match of all the exemplars that make it up. One of the most influential exemplar models has been the context model developed by Medin and Schaffer (1978) and extended by Nosofsky (1986) and Estes (1986).

The context model is mathematically equivalent to the FLMP if each category is represented by one exemplar. The reason is that the context model and the FLMP contain essentially identical evaluation, integration, and decision operations. In the context model, a test stimulus acts as a retrieval cue for exemplar representations in memory. Exemplars in the context model are represented by a set of attributes. Evaluation produces a goodness-of-match value of each attribute of the test item with the corresponding attribute of each exemplar of each category. This goodness of match is represented by a value between 0 and 1, corresponding to the similarity of an attribute of the test item and the corresponding representation of the attribute in memory. The integration of the goodness-of-match values across the different attributes is assumed to be multiplicative (as it is in the FLMP). Finally, decision is accomplished via the RGR in the same manner as in the FLMP. The two models make equivalent predictions in the case in which only one exemplar is assumed in the context model, and the representation of the exemplar is equivalent to the summary description in the FLMP.

Medin and Schaffer (1978) also pointed out the value of multiplicative relative to additive integration in their description of the combination of dimensions to determine overall similarity. Given a multiplicative integration, the overall similarity of a yellow circle and a blue triangle would not be much less than the overall similarity of a yellow circle and a yellow triangle because similarity along color would have very little influence on performance given the gross mismatch on shape. Given additive integration, on the other hand, the overall similarity of a yellow circle and a blue triangle would be significantly less than the overall similarity of a yellow circle and a yellow triangle because similarity of color would add to the goodness of match regardless of the gross mismatch on shape. In a multiplicative combination rule, a single dimension of difference can overrule several dimensions of sameness.

Summary-description models are easily extended to include multiple descriptions of a given category. The most natural extension is to use the summary description in memory that gives the best match with the test item. In this case, integration would involve the goodness of match of the best fitting exemplar of the category of interest. This computation corresponds to the computation of disjunction. Given a definition of conjunction, disjunction can be computed with DeMorgan's law. Given two exemplars E_{k1} and E_{k2} making up the description of category k, the goodness-of-match a_k of a test item with category k would be given by the disjunction of the goodness-of-match values of the test item with the exemplars E_{k1} and E_{k2}:

$$a_k = t(E_{k1} \text{ or } E_{k2}) = t(E_{k2}) + t(E_{k2}) - t(E_{k1}) \times t(E_{k2}). \quad (19)$$

With this definition of disjunction, the context model and the FLMP with multiple summary descriptions are no longer mathematically equivalent. Consider a situation with two contrasting categories with two exemplars in each category. Define a_{kj} as the goodness of match of a test item with exemplar j from category k. Thus, a_{11} and a_{12} are the goodness of match of the test item with Exemplars 1 and 2 from Category 1. The support for Category 1, notated a_1, would be given by

$$a_1 = a_{11} + a_{12} - (a_{11} \times a_{12}).$$

Analogously, the support for category 2 is given by

$$a_2 = a_{21} + a_{22} - (a_{21} \times a_{22}).$$

In the context model, on the other hand, the degree of support for a given category is the simple sum of the degree of support of all exemplars within that category:

$$a_1 = a_{11} + a_{12}$$

and

$$a_2 = a_{21} + a_{22}.$$

At a quantitative level, the models differ on how all of the exemplars in memory contribute to the overall goodness of match of a test stimulus to a category. Thus, in principle, this extension of the FLMP could be tested against the context model. To do so, however, lies outside the scope of this article.

Estes (1986) has also shown a close correspondence between exemplar and summary-description models. Estes (1986) did not address the integration question directly but concentrated instead on the retrieval of exemplar representations, the use of

feature and pattern frequencies, and the existence of prototypes. In addition, we have not addressed the learning of categories. Perhaps people use exemplar-based categorization early in learning before a reliable summary description is developed (Estes, 1986).

Discussion

Previous Rejections of Optimal Behavior

Our analyses in the Empirical Predictions and Tests of the Models section provide preliminary support for the FLMP, an optimal model of pattern recognition. There is also a history of study of the psychological validity of normative (optimal) models in decision making and judgment (Anderson & Shanteau, 1970; Arkes & Hammond, 1986). In contrast to our conclusions, the consensus from the research is that normative models are invalid. Previous research has rejected Bayes's theorem in various judgmental situations (Kahneman & Tversky, 1972). As an example, tests of Bayes's theorem have required estimates of probability in some variant of the two-urn task (Slovic & Lichtenstein, 1971). Subjects see two urns and are told the proportion of red and blue beads in each urn. One urn is picked with some probability, and a sample of beads is drawn. Given the sample, the subject estimates which urn was, in fact, picked. The probability of picking an urn, the relative proportion of beads in each urn, the sample size, and the sample makeup can be varied. The typical result is that subjects behave less optimally than predicted by Bayes's theorem (e.g., Leon & Anderson, 1974).

Our impression is that the rejections of the Bayesian model have been premature. The rejection of Bayes's theorem in many experiments has been a rejection of the normative form of the model rather than a psychological form of the model. Predictions have been derived on the basis of the objective rather than the subjective sources of information. Our implementations of the models, on the other hand, allow for subjective values for the various objective sources of information. Consider a test of the Bayesian model in situations in which subjective base rates are assumed to be equal to objective base rates. In these cases, performance falls short of the predictions of Bayes's theorem (Leon & Anderson, 1974). Central to the current theoretical framework, however, is the evaluation stage that transforms the objective source of information into some subjective value. Thus, performance could still fall short of the optimally objective prediction but might still be described by the same optimal algorithm if subjective values are assumed.

Given the mathematical correspondence between Bayes's theorem and the FLMP, the question arises whether one can be justified over the other. Deciding between the models boils down to beliefs about the psychological reality of the currency assumed by the models. For Bayes's theorem, the currency is probability; for the FLMP, it is truth value. Traditionally, the use of probabilities in psychology has been associated with threshold or categorical models (Massaro, 1975). Thus, the use of fuzzy-truth values represents a shift away from these models to continuously valued states of information.

Bayes's theorem could easily be interpreted as the subject having only categorical information about a given hypothesis (response alternative). Research has proved, however, that people have information about the goodness of match of an instance with a category (Rosch, 1975). As an example, a sparrow provides a better match to the concept of bird than does a penguin. Within a model based on Bayes's theorem, the probability of bird given sparrow would have to be greater than the probability of bird given penguin. With probability interpreted as relative frequency, the difference would imply that the proportion of sparrows that are birds is greater than the proportion of penguins that are birds. However, this difference in probability is not what is meant when people say that a sparrow is a better bird than is a penguin. Differences in truth value appear to capture the difference between penguin and sparrow more accurately. The proposition that a sparrow is a bird is more true than the proposition that a penguin is a bird. The representation of birdness in terms of truth values appears more reasonable than a representation in terms of probabilities.

Relationship Between Luce's Choice Rule and Thurstone's Case V

The RGR and the CR encompass significant aspects of Luce's (1959, 1977) choice axiom and Thurstone's (1927) theory of comparative judgment, respectively. In the choice axiom, the choice objects are represented by scale values (analogous to the discriminal processes of Case V of Thurstone). The choice axiom holds if and only if (a) the RGR holds, (b) the scale value representing an object does not change with changes in the response alternatives used in the choice task, and (c) the response alternatives defined as *irrelevant* do not enter into the RGR. Mathematical psychologists have been aware of a close relationship between Thurstone's theory of comparative judgment and Luce's (1959) choice axiom since the latter's development. Luce (1959) proved that the choice axiom is equivalent to a version of Thurstone's theory in which the differences between the discriminal processes have a logistic distribution instead of the normal distribution implied by Case V (Adams & Messick, 1957). There is equivalence between the two models if and only if the differences between the discriminal processes are logistic random variables.

Yellott (1977) observed that knowing the distribution of the discriminal processes themselves, not simply the distribution of the differences, is important. In addition, can the relationship between the two models be generalized to sets of alternatives greater than two? If the discriminal processes are assumed to have the double-exponential distribution, then the differences will be logistic, and the two models are equivalent for any choice experiment, not simply for pair comparisons (Yellott, 1977). Also, for pair comparisons, distributions other than the double-exponential type yield equivalence between the two models. For three or more alternatives, however, the double-exponential distribution is unique.

Information Manipulation Versus Use

The approach that we have taken in this article involves the systematic manipulation of the properties of patterns. Subjects identify patterns modified in systematic ways, and their responses are used to test quantitative models of the identification

process (Naus & Shillman, 1976; Oden, 1979). An important distinction must be made between the stimulus characteristics of the patterns that are manipulated in the experiment and the features that the perceiver actually uses in the identification of the patterns (Massaro & Schmuller, 1975, p. 209). Patterns can be described by an almost endless number of characteristics or properties (Palmer, 1978), and only a small set of these will be psychologically real. Thus, manipulation of a particular characteristic does not ensure that it is a feature that is used in pattern recognition (Cheng & Pachella, 1984; Sattath & Tversky, 1987). Estes (1986) observed that the researcher needs to know the sources of information actually being used by the subjects in order to provide valid tests of models of categorization. Which characteristics function as features remains a psychological question to be answered.

The paradigm that we have proposed, however, also allows the experimenter to test which sources of information are being used by the perceiver. In speech perception, a voicing distinction allows us to perceive a difference between the verb in the phrase "to use" and the noun in the phrase "the use." Speech scientists believed that consonant duration relative to vowel duration (called the *C/V ratio*) was the critical cue to the voicing judgments (Denes, 1955; Port & Dalby, 1982). However, Massaro and Cohen (1977, 1983) showed that this cue is invalid, when the results are analyzed in the manner developed in this article. A model based on C/V ratio gives a much poorer description of existing results than does a model based on the assumption of independent consonant and vowel duration cues (Derr & Massaro, 1980). Thus, the research strategy developed here not only addresses how different sources of information are evaluated and integrated, it can uncover what sources of information are actually used. We believe that the research paradigm confronts both the important psychophysical question of the nature of information and the process question of how the information is transformed and mapped into behavior.

Equivalence of Models Under Currency Transformation

The distinctions we have drawn between the various integration models rely on the assumption that the psychological values can be measured on valid interval scales. Some psychologists are unwilling to accept this assumption and believe that only ordinal data are meaningful (Krantz, Luce, Suppes, & Tversky, 1971). If the currency or scale values are defined up to only a monotone transformation, then one cannot distinguish among the different integration functions, and indeed, the integration function can always be taken to be summation. For example, a logarithmic transformation applied to the evaluation outputs of the FLMP can be integrated by summation rather than multiplication, and then transformed back by an exponential transformation before the RGR decision process, to obtain an additive integration model that is observationally equivalent to the FLMP. Because transformations of a similar sort can be made for any of the models, the integration function is not unique when arbitrary transformations are permitted.

Our position is that implementable models must specify transformations of the currency as part of the evaluation, integration, and decision processes; that is, they must be psychologically motivated. If these processes are only defined ordinally, then there are infinitely many degrees of freedom because the space of monotone transformations is infinitely dimensional. Such excessive lack of parsimony precludes most meaningful empirical comparisons.

The quantitative models described in Table 1 were all motivated by the underlying psychological processes assumed by the models. For example, the z transformation in the TSD model is based on the assumption of normal noise. This assumption is not only psychologically plausible, it can be tested against empirical data. The particular transformations and integration functions we have developed and tested, of course, are not unique. For example, we discovered that the same TSD model results from (a) an additive integration function applied to noise-free inputs followed by a noisy criterion rule (as in LIM–CR) or (b) a Euclidean distance integration function applied to z-transformed inputs (as in MDS). The point is that some specific transformations must be assumed to compare the models empirically, and we have sought the simplest and most natural specification for each model.

Summary

Our main analytical results are as follows. (a) The FLMP, with truth values estimated from the data, is observationally equivalent as a model of information integration to an optimal model with Bayesian integration and subjective probabilities estimated from the data. (b) A two-level connectionist model (the CMP) is mathematically equivalent to the FLMP for experiments with two response alternatives. Experiments with three or more response alternatives can distinguish between these two models. In this case, the CMP model is prescriptively inferior (i.e., nonoptimal) and descriptively inadequate. (c) A LIM–RGR predicts additive results in both categorical-response and graded-response experiments. This prediction is not only nonoptimal, it gives a poor description of actual behavior. (d) The TSD and a LIM–CR are observationally equivalent in two-alternative categorical response experiments. TSD and LIM–CR are not observationally equivalent to the FLMP but are optimal only under the further restrictions that the information-evaluation process involves constant-variance normal noise. Multidimensional scaling (MDS) can be formalized to mimic either TSD or the FLMP, depending on the assumptions that are made about noise and the distance-integration function.

The point of our analytical and descriptive exercises is to lay the foundation for valid experimental tests of the models. It should now be clear that several experimental tasks are incapable of distinguishing among some of these models. On the other hand, factorial and expanded factorial experiments with four response alternatives can distinguish among the models. We were also successful in distinguishing among the models' descriptions of actual empirical results. The experimental tasks included factorial designs with two categorical responses, four categorical responses, and graded responses. This analysis suggested that the FLMP is not only optimal but provides an adequate account of performance in all of these tasks.

We caution to add that predictive superiority in one experimental domain (e.g., G–Q pattern recognition) does not necessarily imply superiority in another domain (e.g., judgment). Also, predictive power is not the only criterion by which psy-

chologists choose models; certainly, conceptual appeal also matters. Nevertheless, we trust that our analysis will help guide the assessment of the information-integration models we have presented and ultimately encourage the formulation of more refined models.

References

Adams, E., & Messick, S. (1957). *The axiomization of Thurstone's successive intervals and paired comparison scaling models* (Tech. Rep. No. 12). Stanford, CA: Applied Mathematics and Statistics Laboratory, Stanford University.

Anderson, N. H. (1962). Application of an additive model to impression formation. *Science, 138,* 817–818.

Anderson, N. H. (1965). Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology, 70,* 394–400.

Anderson, N. H. (1970). Functional measurement and psychophysical judgment. *Psychological Review, 77,* 153–170.

Anderson, N. H. (1973). Functional measurement of social desirability. *Sociometry, 36,* 89–98.

Anderson, N. H. (1974). Algebraic models in perception. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception: Vol. 2. Psychophysical judgement and measurement.* New York: Academic Press.

Anderson, N. H. (1981). *Foundations of information integration theory.* New York: Academic Press.

Anderson, N. H. (1982). *Methods of information integration theory.* New York: Academic Press.

Anderson, N. H., & Cuneo, D. O. (1978). The height + width rule in children's judgments of quantity. *Journal of Experimental Psychology: General, 107,* 335–378.

Anderson, N. H., & Shanteau, J. C. (1970). Information integration in risky decision making. *Journal of Experimental Psychology, 84,* 441–451.

Arkes, H. R., & Hammond, K. R. (Eds.). (1986). *Judgment and decision making: An interdisciplinary reader.* New York: Cambridge University Press.

Asch, S. E. (1964). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41,* 258–290.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 33–53.

Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review, 95,* 124–150.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review, 93,* 154–179.

Bouma, H. (1971). Visual recognition of isolated lower-case letters. *Vision Research, 11,* 459–474.

Braida, L. D., & Durlach, N. I. (1972). Intensity perception II: Resolution in one-interval paradigms. *Journal of the Acoustical Society of America, 51,* 483–502.

Bruno, N., & Cutting, J. E. (1988). Minimodularity and the perception of layout. *Journal of Experimental Psychology: General, 117,* 161–170.

Carterette, E. C., Friedman, M. P., & Wyman, M. J. (1966). Feedback and psychophysical variables in signal detection. *Journal of Acoustical Society of America, 39,* 1051–1055.

Chandler, J. P. (1969). Subroutine STEPIT—Finds local minima of a smooth function of several parameters. *Behavioral Science, 14,* 81–82.

Cheng, P. W., & Pachella, R. G. (1984). A psychophysical approach to dimensional separability. *Cognitive Psychology, 16,* 279–304.

Clarke, F. R. (1957). Constant-ratio rule for confusion matrices in speech communication. *Journal of the Acoustical Society of America, 29,* 715–720.

Davison, M., & McCarthy, D. (1988). *The matching law: A research review.* Hillsdale, NJ: Erlbaum.

DeGroot, M. H. (1970). *Optimal statistical decisions.* New York: McGraw-Hill.

Denes, P. (1955). Effect of duration on the perception of voicing. *Journal of the Acoustical Society of America, 27,* 761–764.

Derr, M. A., & Massaro, D. M. (1980). The contribution of vowel duration, Fo contour, and frication duration as cues to the /juz/-jus/ distinction. *Perception & Psychophysics, 27,* 51–59.

Dosher, B. A., Sperling, G., & Wurst, S. A. (1986). Tradeoffs between stereopsis and proximity luminance covariance as determinants of perceived 3D structure. *Vision Research, 26,* 973–990.

Estes, W. K. (1984). Global and local control of choice behavior by cyclically varying outcome probabilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 258–270.

Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology, 18,* 500–549.

Falmagne, J.-C. (1985). *Elements of psychophysical theory.* New York: Oxford University Press.

Fidell, S. (1970). Sensory function in multimodal signal detection. *Journal of the Acoustical Society of America, 47,* 1009–1015.

Fisher, R. A. (1935). *The design of experiments.* Edinburgh, Scotland: Oliver & Boyd.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Erlbaum.

Gilmore, G. C., Hersh, H., Caramazza, A., & Griffin, J. (1979). Multidimensional letter similarity derived from recognition errors. *Perception & Psychophysics, 25,* 425–431.

Goguen, J. A. (1969). The logic of inexact concepts. *Synthese, 19,* 325–373.

Golden, R. M. (1988). A unified framework for connectionist models. *Biological Cybernetics, 59,* 109–120.

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

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3,* 430–454.

Kornbrot, D. E. (1978). Theoretical and empirical comparison of Luce's choice model and logistic Thurstone model of categorical judgment. *Perception & Psychophysics, 24,* 193–208.

Krantz, D. H. (1969). Threshold theories of signal detection. *Psychological Review, 76,* 308–324.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement* (Vol. 1). New York: Academic Press.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29,* 1–27.

Leon, M., & Anderson, N. H. (1974). A ratio rule from integration theory applied to inference judgments. *Journal of Experimental Psychology, 102,* 27–36.

Loomis, J. M. (1982). Analysis of tactile and visual confusion matrices. *Perception & Psychophysics, 31,* 41–52.

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

Luce, R. D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology, 15,* 215–233.

Massaro, D. W. (1969). The role of the decision system in sensory and memory experiments using confidence judgments. *Perception & Psychophysics, 5,* 270–272.

Massaro, D. W. (1975). *Experimental psychology and information processing.* Chicago: Rand McNally.

Massaro, D. W. (1984). Building and testing models of reading processes. In P. D. Pearson (Ed.), *Handbook of reading research* (pp. 111–146). New York: Longman.

Massaro, D. W. (1985). Attention and perception: An information-integration perspective. *Acta Psychologica, 60,* 211–243.

Massaro, D. W. (1987a). Integrating multiple sources of information in listening and reading. In D. A. Allport, D. G. MacKay, W. Prinz, & E. Scheerer (Eds.), *Language perception and production: Shared mechanisms in listening, speaking, reading and writing* (pp. 111–129). London: Academic Press.

Massaro, D. W. (1987b). *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, NJ: Erlbaum.

Massaro, D. W. (1988a). Ambiguity in perception and experimentation. *Journal of Experimental Psychology, 117,* 417–421.

Massaro, D. W. (1988b). Some criticisms of connectionist models of human performance. *Journal of Memory and Language, 27,* 213–234.

Massaro, D. W., & Cohen, M. M. (1977). Voice onset time and fundamental frequency as cues to the /zi/-/si/ distinction. *Perception & Psychophysics, 22,* 373–382.

Massaro, D. W., & Cohen, M. M. (1983). Consonant/vowel ratio: An improbable cue in speech. *Perception & Psychophysics, 33,* 501–505.

Massaro, D. W., & Cohen, M. M. (1987). Process and connectionist models of pattern recognition. *Proceedings of the Ninth Annual Conference of the Cognitive Science Society* (pp. 258–264). Hillsdale, NJ: Erlbaum.

Massaro, D. W., & Cohen, M. M. (1988). *Perceiving single or multiple objects.* Unpublished manuscript.

Massaro, D. W., & Hary, J. M. (1986). Addressing issues in letter recognition. *Psychological Research, 48,* 123–132.

Massaro, D. W., & Schmuller, J. (1975). Visual features, preperceptual storage, and processing time in reading. In D. W. Massaro (Ed.), *Understanding language: An information-processing analysis of speech perception, reading, and psycholinguistics* (pp. 207–240). New York: Academic Press.

Massaro, D. W., Tseng, C., & Cohen, M. M. (1983). Vowel and lexical tone perception in Mandarin Chinese: Psycholinguistic and psychoacoustic contributions. *Quantitative Linguistics, 19,* 76–102.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1–86.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238.

Minsky, M. L., & Papert, S. A. (1988, expanded ed.). *Perceptrons.* Cambridge, MA: MIT Press. (Originally published 1969)

Myers, J. L. (1976). Probability learning and sequence learning. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes* (pp. 171–205). Hillsdale, NJ: Erlbaum.

Naus, M. J., & Shillman, R. J. (1976). Why a Y is not a V: A new look at the distinctive features of letters. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 394–400.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57.

Oden, G. C. (1977). Fuzziness in semantic memory: Choosing exemplars of subjective categories. *Memory & Cognition, 5,* 198–204.

Oden, G. C. (1979). A fuzzy logical model of letter identification. *Journal of Experimental Psychology: Human Perception and Performance, 5,* 336–352.

Oden, G. C. (1981). A fuzzy propositional model of concept structure and use: A case study in object identification. In G. W. Lasker (Ed.), *Applied systems and cybernetics* (Vol. 6, pp. 2890–2897). Elmsford, NY: Pergamon Press.

Oden, G. C. (1984). Dependence, independence, and emergence of word features. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 394–405.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review, 85,* 172–191.

Palmer, S. E. (1978). Fundamental aspects of cognitive representation. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 574–590). Hillsdale, NJ: Erlbaum.

Perkell, J. S., & Klatt, D. H. (Eds.). (1986). *Invariance and variability in speech processes.* Hillsdale, NJ: Erlbaum.

Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *Transactions of IRE Professional Group on Information Theory, PGIT-4,* 171–212.

Port, R. F., & Dalby, J. (1982). Consonant/vowel ratio as a cue for voicing in English. *Perception & Psychophysics, 32,* 141–152.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104,* 192–233.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65,* 386–408.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Vol. 1. Foundations.* Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (Eds.). (1986). *Parallel distributed processing: Vol. 1. Foundations.* Cambridge, MA: MIT Press.

Sattath, S., & Tversky, A. (1987). On the relation between common and distinctive feature models. *Psychological Review, 94,* 16–22.

Selfridge, O. G. (1959). Pandemonium: A paradigm for learning. In *Mechanisms of thought processes* (pp. 511–526). London: Her Majesty's Stationery Office.

Shaw, M. L. (1982). Attending to multiple sources of information: I. The integration of information in decision making. *Cognitive Psychology, 14,* 353–409.

Shaw, M. L., Mulligan, R. M., & Stone, L. D. (1983). Two-state versus continuous-state stimulus representations: A test based on attentional constraints. *Perception & Psychophysics, 33,* 338–354.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika, 22,* 325–345.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika, 27,* 125–140.

Shepard, R. N. (1986). Discrimination and generalization in identification and classification: Comment on Nosofsky. *Journal of Experimental Psychology: General, 115,* 58–61.

Shepard, R. N. (1988). George Miller's data and the development of methods for representing cognitive structures. In W. Hirst (Ed.), *The making of cognitive science: Essays in honor of George A. Miller* (pp. 45–70). Cambridge, England: Cambridge University Press.

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance, 6,* 649–744.

Stanislaw, H. (1988). Methodological considerations of the study of multimodal signal detection. *Perception & Psychophysics, 44,* 541–550.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica, 30,* 276–315.

Stevens, S. S. (1961). Is there a quantal threshold? In W. A. Rosenblith (Ed.), *Sensory communication.* New York: Technology Press & Wiley.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Cambridge, MA: Belknap Press.

Swets, J. A. (1961). Is there a sensory threshold? *Science, 134,* 168–177.

Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review, 68,* 301–340.

Tanner, W. P., Jr., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review, 61,* 401–409.

Thomas, E. A. C., & Legge, D. (1970). Probability matching as a basis

for detection and recognition decisions. *Psychological Review, 77,* 65–72.

Thurstone, L. L. (1927). Psychophysical analysis. *American Journal of Psychology, 38,* 368–389.

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

Townsend, J. T. (1984). Uncovering mental processes with factorial experiments. *Journal of Mathematical Psychology, 28,* 363–400.

Townsend, J. T., Hu, G. G., & Kadlec, H. (1988). Feature sensitivity, bias, and interdependencies as a function of energy and payoffs. *Perception & Psychophysics, 43,* 575–591.

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

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88,* 638–667.

Wickelgren, W. A. (1968). Testing two-state theories with operating characteristics and a posteriori probabilities. *Psychological Bulletin, 69,* 126–131.

Yellott, J. I., Jr. (1977). The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology, 15,* 109–144.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8,* 338–353.

Appendix

Description of Notation Used in the Analysis of Tasks and Models

Notation	Description	Notation	Description
$X, Y, \cdots$	Sources of information	S_i	Test stimulus i
$X_i, Y_j, \cdots$	ith, jth, $\cdots$ levels of $X, Y, \cdots$	d'	d' of theory of signal detectability
$x, y, \cdots$	Scale values given by evaluation of $X, Y, \cdots$	d'_C	d' given the source of information C
A_k	Response alternative k	d'_{CO}	d' given the two sources of information C and O
a_k	Scale value given by integration, support for A_k	E_{kj}	Exemplar j from category k
C	Closed property of oval of G–Q test letters	a_{kj}	Goodness of match of test item with exemplar j from category k
c	Evaluation of C	RGR	Relative goodness rule
O	Oblique property of line of G–Q test letters	CR	Criterion rule
o	Evaluation of O	FLMP	Fuzzy-logical model of perception
L	Degree of presence of line in G–Q–C–O test letters	TSD	Theory of signal detectability
l	Evaluation of L	LIM	Linear integration model
$Pr(A_k \mid X = X_i, Y = Y_j)$	Probability of k response given the ith level of X and the jth level of Y. Also written as $P(k \mid X_i Y_j)$ or $P(k \mid XY)$	CMP	Connectionist model of perception
		MDS	Multidimensional scaling

Received June 13, 1988
Revision received August 17, 1989
Accepted August 31, 1989 ■