rborough (Eds.), Cognition
Cambridge: MIT Press 1998

Chapter 8

Models for Reading Letters and Words

Dominic W. Massaro

Editors' Introduction

In this chapter, Dominic Massaro gives us a case study of visual recognition of letters and words. Recognizing letters of the alphabet may seem like a trivial and mundane task. In fact, simple to be interesting. But though people can do it with ease, it is important to remember that only now, after decades of work, are computers beginning to show some success in this task. And, in fact, it takes children many years of practice. Apparently, it is a lot harder than it looks, and the ease with which people can do it is a testimony to the sophistication of our visual pattern-recognition skills.

Massaro takes us on a tour through several different theoretical accounts of how people manage to recognize letters and words. At each step, he shows us that by clearly specifying a theory, we are in a much better position to evaluate its strengths and weaknesses. Using some of his own research, he demonstrates the failures of a number of perceptual theories, including some recent work in the "hot" area of artificial neural networks. He shows that simple letter recognition requires complex perceptual abilities to evaluate and integrate the characteristics or features of the letters, and that this analysis must then be followed by a decision-making process. In this respect, his treatment converges with the general account of perception and decision making described by Swets in chapter 13. Massaro argues that the most successful theoretical approach must be based on "fuzzy logic," an approach that has recently been used in commercial applications.

Chapter Contents

8.1 Introduction

A distinguished singer and music teacher found himself unable to recognize the simplest and most familiar patterns. Here was a man of cultivation, imagination, and charm who had no noticeable deficit. However, having decided to leave a meeting, he looked around for his hat, reached out and took hold of his wife's head and tried to put it on. Mistaking his wife's head for a hat was pathetic enough to inspire Oliver Sacks to entitle his engaging anthology of clinical neurology dysfunctions *The Man Who Mistook His Wife for a Hat and Other Clinical Tales* (1970).

What critical ingredient could this poor gentleman have lost from his mind's ability to recognize and impose order on the blooming, buzzing confusion around us? This chapter cannot answer this question, but the goal is to illustrate how developing and testing computational models can inform us about the psychological processes involved in perceiving, recognizing, and categorizing the world. To simplify the discussion, I will focus on the question of how we perceive letters and words.

When faced with a written word we seem to have no choice but to read it. Our phenomenal experience attests to this fact, as do experiments, demonstrations of the Stroop effect (1935). Take a look at the colored patterns [illegible]

Figure 8.1
Be cool. The same visual configuration can mean different things in different contexts.

the word green in red ink, and so on. Read the list of words aloud from top to bottom. This task is not meant to insult your intelligence, but to serve as a baseline for the next task. Now name the colors of the words from top to bottom. You will experience firsthand that having the colors presented in incompatible written color names interferes with naming the colors. Although your intention is to name the colors and ignore the words, it is not possible. Reading words is such an overlearned skill, it is not easily put on hold.

Achieving this level of reading skill takes time. A millionaire recently admitted that he was illiterate. What is impressive beyond his success at deception through school and college is the time he required to learn to read at the age of 60. Learning to read competently involved sixty 40-hour weeks of studying and sounding out words. This extended period of study might seem excessive, but probably is in the ballpark of the time most of us required to learn to read. Clearly, the expertise of an adult reader is based on extensive practice. As Huey (1908) noted in 1908, reading is a remarkably complex skill, and I will address just one small aspect having to do with how we recognize letters and words.

In addition to being experts in letter recognition, we are especially good at recognizing them when they spell words. Our knowledge of spelling and the context provided by the other letters of a word help us recognize individual letters within words. Figure 8.1 illustrates this idea by demonstrating how two identical visual patterns can be interpreted as different letters. Although the last letter of the first word is visually the same as the first letter of the second word, what we know about English spelling demands that they be interpreted differently. We would expect that such knowledge would enable us to identify words even when the visual information is incomplete or fuzzy, or to extract meaning from a page of text without analyzing all the visual information [illegible]

8.2 Pattern Recognition

8.2.1 Domains of Pattern Recognition

The anecdote about the man who mistook his wife for a hat illustrates a failure of pattern recognition. I use the term *pattern recognition* to describe what is commonly meant by perception, recognition, identification, and categorization. Although these terms have different meanings, they are all concerned with roughly the same phenomenon. *Perception* describes our awareness of some environmental event. *Recognition* means recognizing something we experienced previously. *Identification* involves making a unique response to each unique stimulus. *Categorization* means placing several noticeably different stimuli into the same category. For example, a child sees a dog, recognizes it as a dog she has seen before, identifies it as Fido, and categorizes it as a dog. Recognition, identification, and categorization appear to be central to perceptual and cognitive functioning, and they appear to entail the same fundamental processes. Pattern recognition is fundamental in such different domains as playing chess, examining X-rays (see Swets, chap. 13, this volume), and reading.

8.2.2 Computational Models

This chapter promotes the value of computational models in the enterprise of understanding how humans accomplish feats such as pattern recognition. Although each of us probably has some understanding of each of the two words in *computational model*, it is worthwhile to describe my use of this topic. A *model* is a specific form of a theory. *Computational* usually refers to systems that perform specific operations or computations, such as computers. However, a computational model is not necessarily a computer model but might consist of a set of operations that are described in qualitative form, which is the case for most of the models discussed in this chapter. The value of a computational model is that it can make precise predictions that can be compared with human performance.

8.3 Approaches to Pattern Recognition

8.3.1 Template Matching

You might have wondered why the digits in account numbers on personal checks have unique and distinctive shapes. The reason is that these numbers are recognized by machines that use *template matching*, a form of holistic recognition in which the units of analysis are the same size as the patterns to be recognized. Basically, the machine has a template for each digit, somewhat analogous to a paper cutout of the digit, and recognition is based on the best-matching template. Unique and distinctive shapes for the digits make recognition much easier.

Similarly, a child's toy robot that can recognize a few voice commands also uses template matching. This toy has a learning mode and a performance mode. In the learning mode, the robot stores specific voice commands for moving forward, backing up, turning left or right, stopping, greeting, and lifting up or down. The commands can be words or short phrases, such as "Move Forward", "Back up", "Turn right", "Left turn", "Stop", "Hello", "Lift it up", and "Put them down". Similar to the digit templates in the account number example, the robot stores a template of the sound of each of the commands specified by the teacher. (Notice that I chose the commands to be very different from one another; distinctive commands make the task easier.) In its performance mode, the speaker utters a command, and the robot matches the auditory input with each of the templates in memory. The robot's response to the command is determined by the template that gives the best match.

We might expect that some characteristics or features of a pattern would be more relevant for recognition than others. In English, for example, the pitch of a speaker's voice is not important in identifying a spoken word. However, the robot performs best when commands are given by the original teacher speaking at the same rate. Changing properties that do not affect the meaning of the command, such as voice pitch, can reduce performance considerably. These limitations arise because the template matching routine has no method for ignoring less relevant features and utilizing the more relevant features. The toy robot does not know that we speak at different rates, with different voices, and with different dialects. All of these differences will contribute to a mismatch between the specific templates stored in memory and a new command.

A related limitation is that simple template matching has no way to evaluate the importance of mismatches. Consider a case in which there is only partial information about the stimulus. As shown in figure 8.2, about one third of the letters H and N can be eliminated either by removing the middle line in both of the letters or by making the letters out of dashed line segments. However, the same absolute amount of distortion can produce dramatically different results. We cannot recognize which letter is which in the first case, but we can in the second. Eliminating the middle of the letters changes the nature of the patterns, whereas making the continuous lines into dashes does not, and we know enough to dismiss missing information when the nature of the pattern is not changed. Simple template matching routines cannot handle this problem.

The problems with template matching become unwieldy when there are many patterns to be recognized. The toy robot does an acceptable job with eight very different commands, but would have difficulty with a few

Figure 8.2
The letters N and H presented with (a) intact form, (b) the middle line missing, and (c) with the same amount of visual information missing as in (b) but distributed evenly across the letter.

hundred. A child, on the other hand, understands (but does not necessarily obey) an unlimited number of commands despite tremendous variation across different speakers. And readers can recognize a variety of type fonts. Figure 8.3 gives a sample of the many different characters that can be easily recognized. Increasing the number of templates makes the matching process much more difficult because patterns tend to be similar to one another and because the number of template comparisons becomes very large. Given the limitations of template models, psychologists have turned to feature analysis as an alternative description of pattern recognition in humans.

8.3.2 Feature Analysis

In *feature analysis*, each pattern is represented in memory in terms of its component parts and their relationships. For the robot, the command "Go-forward" might be described in terms of the words or syllables that make it up, and the sequential order of these smaller components. Recognition would now involve matching those component parts or features to the spoken command. As I caution my students, however, because matching individual features is the same as matching a set of mini-templates, feature analysis is not a foolproof panacea for the difficulties faced by template matching. The advantage gained by feature analysis is simply in terms of the smaller size of the units being matched. We will see that feature analysis, although also susceptible to error, provides a more parsimonious and [illegible]

Figure 8.3
We are good at recognizing letters in different sizes, different orientations, different type fonts, and even distorted or incomplete letters.

The feature analysis approach has been influenced by work in artificial intelligence, linguistics, and visual neurophysiology. An example of the influence of artificial intelligence is found in Oliver Selfridge's engaging "Pandemonium" model (1959). Selfridge chose the word "demon" to represent an operation or component within the model, and each demon dealt with some aspect of the recognition process. For letter recognition, Selfridge utilized four types of demons arranged in hierarchical order, as illustrated in figure 8.4. First, image demons transduce the stimulus. In robot vision, for example, image demons might convert light energy into electrical signals representing the light intensity at each small region or pixel of the stimulus. Next, the image demons send this information to the feature demons, which correspond to mini-templates that are matched to the image. In figure 8.4 presentation of an uppercase *A* activates a horizontal and two oblique feature demons. Each feature demon "shouts," that is, sends an output message when it matches its feature. The cognitive letter demons, corresponding to letters of the alphabet, know what features make their up, and they monitor the messages from the feature demons. The letter *A* demon gets messages from three feature demons, but the *H* and *B* demons only get a message from the horizontal-feature demon. A letter demon also sends a message to the extent that its appropriate feature demons are active. The buck stops at the decision demon, the mother of all demons, who monitors the letter demons and decides in favor of the most active demon. Given the [illegible] letters, the decision demon [illegible]

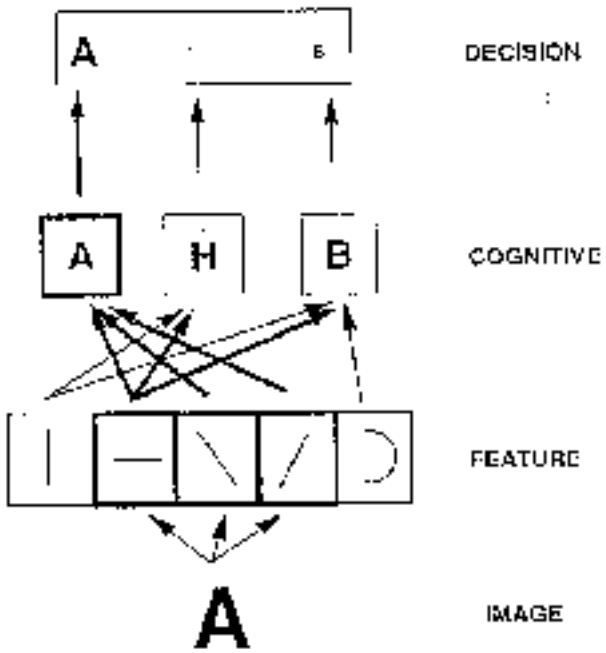

Figure 8.6
The arrangement of the four types of demons for letter recognition in Selfridge's Pandemonium model (1959). The image demons create an image of the input, which is then passed on to the feature demons. Each feature demon becomes active if it sees its own particular feature in the image. Each cognitive demon represents a particular letter. A cognitive demon monitors the feature demons to determine if the features corresponding to its character have been detected. Finally, the decision demon decides which letter is present based on which cognitive demon is most active.

Work in linguistics supports Selfridge's Pandemonium approach. Four decades ago the phoneme was considered to be the minimal speech segment that could change the meaning of a word. For example, the word "bet" has three phonemes /b/, /e/, and /t/. Changing embedded between // refers to phonemes. In English, we change the /t/ to /d/ and get "bed," the /e/ to get "bat," or the /b/ to /p/ to get "pet." Different languages share many of the same phonemes. A family of 50 or 60 phonemes is sufficient to describe all of the world's languages. The phoneme was conceptualized as a fairly fixed template until some linguists came up with the concept of distinctive features, which were properties that distinguished among the different phonemes. Voiced versus voiceless is an example of a distinctive feature, as in the difference between /b/ and /p/. With an unvoiced consonant such as /p/, the consonantal sound is made primarily by passage of air through the mouth.

With a voiced consonant such as /b/, the consonantal sound is accompanied by vibration of the vocal chords or "voice." Thus the two phonemes /p/ and /b/ are similar and differ only in the voicing feature. It turned out that only about a dozen distinctive features were necessary to describe all of the phonemes of a particular language. Each phoneme was simply described in terms of the presence or absence of each of these features, as for example in the voicing of a consonant. For our purposes, the distinctive feature analysis offered an alternative to template matching of phonemes. Speech recognition could be carried out, not by template matching, but by analysis of distinctive features.

Finally, work in visual neurophysiology carried out over fifty years ago provides additional evidence that perception is based on feature analysis. Receptors in the eye transduce light energy into neural responses that pass up the visual pathways to cells in visual cortex. Neural responses recorded from visual cortex in an anesthetized cat show that individual cortical cells act like Selfridge's feature demons and respond selectively to aspects of the visual input. One set of cells became known as "simple cells"; these cells are essentially edge detectors that respond only to a border between light and dark. Other simple cells are line detectors or slit detectors, both of which can be considered modified edge detectors. Simple cells respond only when a stimulus is presented at a specific location and orientation on the retina. Other cells, called "complex cells," are more selective in their response and respond only to lines of a particular width and orientation. Finally, "hypercomplex cells" are even more discriminating in that the lines also must be a specific length. Here before our eyes, or at least the cat's, is an example of feature analysis. This feature analysis also appears to be hierarchical—simple cells send information to complex cells, which send information to hypercomplex cells and so on, eventually leading to a cell so specialized that some people speculated that it would recognize your grandmother.

Although we usually experience a pattern, such as a letter, as an integral whole, the work in artificial intelligence, linguistics, and neurophysiology encouraged psychologists to develop feature models of pattern recognition with the following characteristics. First, patterns are described in terms of their component parts or features, as opposed to holistic templates. Second, these parts or features take on different values for different patterns—the features differentiate the patterns from one another. Thus the letter *A* has oblique lines whereas *H* has vertical ones. Third, the features can be arranged hierarchically, with the smallest features being grouped into somewhat larger features, and the somewhat larger features being grouped into even larger more complex features, and so on. For example, in this hierarchical scheme the combination of two oblique lines in the letter *A* is recognized by an acute angle feature detector that

depends on the output of the oblique line feature detectors. Given this foundation, I now proceed to an analysis of various experimental situations that might allow for empirical tests of models of letter recognition.

8.4 Letter Recognition

Partly motivated by linguistics and by the feature detectors discovered in neurophysiology, Eleanor Gibson (1969) proposed a list of discrete (binary) features for recognizing letters. She assumed that there were detectors in the visual system to recognize these specific features regardless of the length, density, or goodness of the features. A feature detector is discrete or binary if a feature is detected as being either present or absent (or as having one of two values). Gibson's list included a variety of binary feature distinctions such as the presence or absence of a straight vertical line, or of intersecting lines. This scheme was sufficient, in principle, for recognizing the twenty-six uppercase letters of the alphabet.

Gibson's binary features approach assumes that a given feature will activate its detector with the same intensity irrespective of the goodness of that feature. Figure 8.5 shows a set of variations of a given feature that

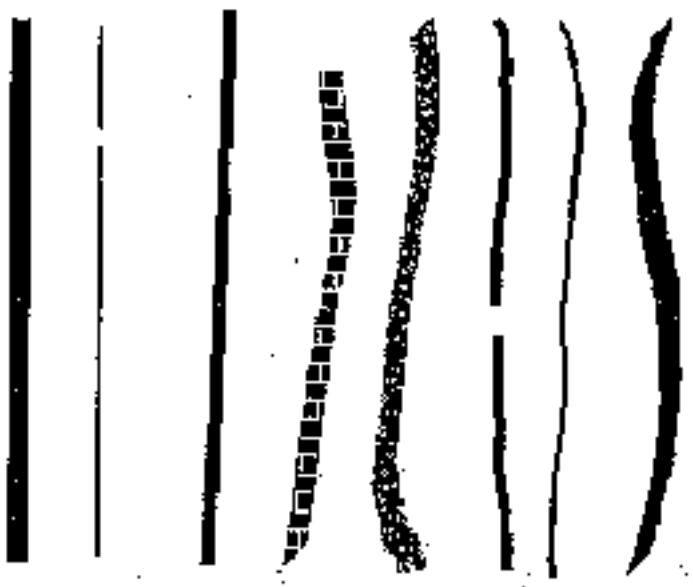

Figure 8.5
A set of variations of a vertical line that might all be equally effective in activating a vertical line feature detector.

would all putatively activate a vertical line feature detector to the same degree. (In this framework, your ability to see differences between the lines depends on other feature detectors in addition to your vertical line feature detector.) I think the discrete feature assumption should be put to experimental test first because it is central to how we think about features. How does one go about testing whether we have feature detectors that convey only discrete (binary or categorical) information versus continuous information (that is, the output of a feature detector varies with the quality of the feature) about each letter feature? Several methodological innovations are necessary to address this question, in addition to the development of computational models.

8.4.1 Fuzzy Letters and Continuous Rating Judgments

Most studies of letter recognition do not address the assumption of discrete features. In a typical experiment, randomly chosen letters are presented for identification. Of course, a typical college student would be perfectly accurate in this task and the results would be uninformative. We have a better chance of testing these ideas if subjects make errors. To induce errors, the test letters can be presented for very short durations and be followed by a visual mask, which is simply another visual display that appears in the same position as the original test letter. A typical result is that letters with more features in common tend to be confused with one another. As an example, *A*, *H*, and *N* are more often mistaken for one another than are *A*, *C*, and *S*. However, these results do not easily test whether perception of the features is continuous or discrete because the results are consistent with either hypothesis.

An innovative approach is to create ambiguous letters by varying the degree to which a feature is present (Oden, 1979). For many pairs of letters, we can create ambiguous letters that resemble both letters. For example, the lowercase letters *c* and *e* appear to differ in the presence of a horizontal line feature for *e* but not *c*. Figure 8.6 shows how I systematically varied the length of this feature to create a continuum of letters differing in terms of how much each letter resembles *e* versus *c*. How readers

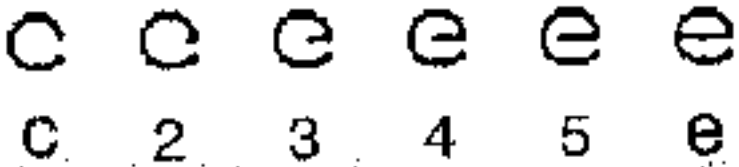

Figure 8.6
Sample letters from a continuum that varies in length of the horizontal line feature.

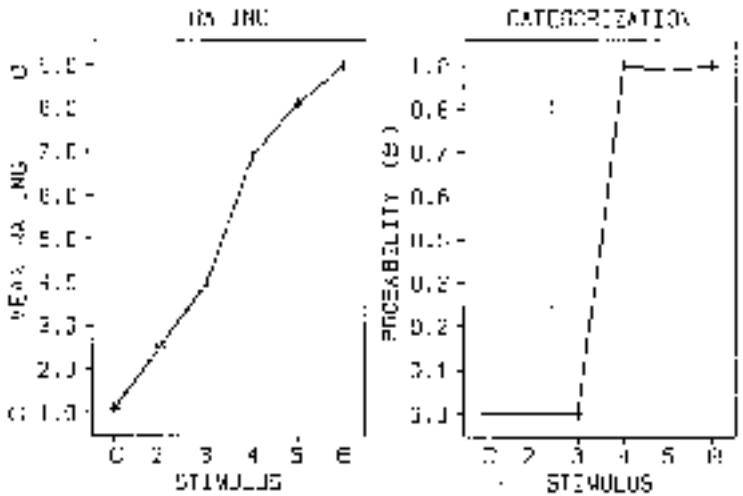

Figure 8.7
Mean (average) ratings on a nine-point scale of the degrees to which a test letter resembles c (1) to e (9) as a function of the six stimulus levels (left panel), and the proportion of e judgments as a function of the six stimulus levels (right panel).

process these ambiguous letters might help us decide whether this feature is perceived discretely or continuously.

In an experiment addressing this issue, each of the six letters shown in figure 8.6 was presented seventy times in random order, and the subject categorized each presentation as an e or a c. The proportions of e responses for each of the stimuli are shown in the right panel of figure 8.7. The results show that letters on the c side of the continuum were consistently categorized as c, and letters on the e side of the continuum were consistently categorized as e. From this it might be (and in fact sometimes has been) concluded that the horizontal line feature is perceived discretely. It was either detected or not (in which case the subject reports e or c respectively), and that all the stimuli within each of the two categories were seen as the same. However, this conclusion would not be justified because the categorization task does not permit subjects to give a direct report of what they saw.

Another possible task is to ask subjects to rate where a presented letter falls on a continuous scale between c and e. What results would you expect from such a rating task relative to the categorization task? Continuous rating judgments may provide a more direct measure of what the observer sees. That is, we might categorize two different animals as dogs even though we see large differences between them. In proposing this, I am assuming that asking for a continuous rating rather than a discrete

categorization judgment does not change the underlying perceptual processes. I think of letter and word recognition as automatic. The nature of the response should not greatly influence the perceptual processing that leads to the response.

In a rating task, the same subject was asked to indicate the point at which each test letter fell on a nine-point scale between c and e. The presentation procedure was the same as in the c-e categorization task. The mean (average) rating judgments are shown in the left panel of figure 8.7. There was roughly a linear (straight-line) relationship between the mean rating and the length of the horizontal line. Contrasting the left and right panels reveals that subjects rated adjacent letters on the continuum as different from one another, even though these were identified equivalently in the c-e categorization task. It apparently follows that subjects probably saw differences among the different letters in the binary choice categorization task even though their c-e categorization judgments do not reflect this fact. But as we shall see, these rating results are also consistent with a model in which the horizontal line feature is perceived discretely as present or absent.

The point I am making here is that categorical (or even continuous) judgments do not directly mirror the processes or the perceptual experience underlying the judgments. As a graduate student, I was struck with this limitation when I was able to make undergraduates call a low tone "high" and a high tone "low" simply on the basis of the feedback they were given after each categorical judgment. Clearly, it would have been a mistake to assume that their perception was accurately mirrored by their categorization. The perceptual report is only one source of evidence among many that the experimenter must use to tap into perceptual processing and to overcome the limitations of any single type of perceptual report.

The results shown in figure 8.7 are not definitive in the sense that they do not answer unambiguously whether letter features are perceived discretely or continuously. In what follows, I show how both discrete and continuous feature recognition theories are consistent with the mean rating results shown in figure 8.7. Then we will see how extending the database and computational modeling can distinguish between the two alternatives. I begin with the discrete model.

8.4.1 Discrete Model

Consider again the stimulus continuum that ranges from c to e. In my formulation of the discrete perceptual model, I assume that there are only two percepts, c or e, that can result from any value along this continuum. But because the same stimulus event does not always lead to the

same response, I assume that this stimulus does not always produce the same percept (possibly because of variations in attention, expectation, etc.). Instead, the percept is probabilistic. The subject sees a *c* with some probability $P_{c|c}$ and an *e* with probability $1 - P_{c|c}$. At stimulus level 1 at the *c* extreme of the continuum, the probability of the *c* percept is very high, so the probability of an *e* percept is very low. As the stimulus level increases, they change the relative likelihood of the two percepts so that *e* is the more likely percept at the *e* end of the stimulus continuum. Thus, in this model, the stimulus is assumed to be perceived as either a *c* or an *e*. The perception is truly discrete, any stimulus presentation being either recognized only as *c* or *e* and nothing in between.

What model discrete perception subjects do when asked to make continuous judgments? They might surreptitiously note the foolishness of the request but attempt to comply by making a range of responses to the same percept. In this case, subjects would choose ratings toward the *c* end of the response scale for the perception of *c* and toward the *e* end for the perception of *e*. Although there would be only two possible percepts, subjects would produce a larger number of different ratings. Also, subjects might not remember exactly how they last rated the *c* and *e* percepts and would therefore give somewhat different ratings on successive trials, again leading to a distribution of rating responses for each of the two percepts.

Given this set of assumptions for the discrete model, an important question is how the mean of the rating responses for each stimulus is expected to change as a function of the position along the stimulus continuum. Consider the *c-e* continuum of six levels illustrated in figure 8.6. We might suppose that the *c* stimulus (stimulus level 1) would reliably lead to a *c* percept, and a subject would respond by giving a rating from 1 to 4 from the *c* end of the response scale. This possibility is represented by the hypothetical data in the bottom panel of figure 8.8. This panel shows the probabilities of ratings of 1, 2, 3, and 4 over many presentations of the same stimulus as approximately 0.13, 0.48, 0.34, and 0.05, respectively, and the mean rating, indicated by the arrow, is about 2.60. On the other hand, presentations of the *e* stimulus should reliably lead to the *e* percept, and a subject should respond with ratings from the *e* end of the response scale, as represented in the top panel of figure 8.6. But an ambiguous stimulus (such as stimulus level 3 or 4) might produce the percept *c* with some probability $P_{c|i}$ and the percept *e* with probability $1 - P_{c|i}$. As a result, the rating response would be selected from the *c* or the *e* end of the response scale on any particular trial depending on how the stimulus was perceived. Finally, a stimulus should be more likely to produce a *c* percept to the extent that it is toward the *c* end of the continuum, as illustrated for stimulus levels 2 to 5 in figure 8.6.

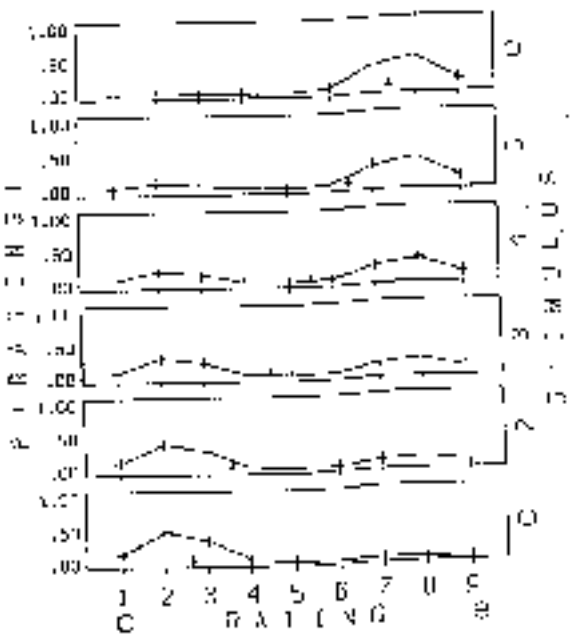

Figure 8.8
Hypothetical distributions of rating responses, according to the discrete model, for each of six stimulus levels along a stimulus continuum. The rating responses are from the left end of the rating scale (i.e., 1–4) when a stimulus is perceived as a *c* (bottom panel) but from the other end of the scale when the stimulus is perceived as an *e* (top panel). For other ambiguous stimuli, the rating responses are drawn from these two distributions. The downward pointing arrows mark the means of the rating responses for each of the six stimulus levels.

From this argument we see that the discrete model predicts that rating responses for an ambiguous stimulus will actually be a mixture of ratings generated by the *c* and *e* percepts. As we move along the stimulus continuum, the proportion of ratings generated from the distribution of *c* ratings will increase with increases in $P_{c|c}$, while the proportion of *e* ratings will decrease. The arrow in each panel of figure 8.8 indicates the mean of the hypothetical rating responses to that stimulus. Replotting these mean ratings, figure 8.9 illustrates that continuous changes in the mean rating response with continuous changes in the stimulus can be predicted by the discrete model. In fact, these predictions of the discrete model are very similar to the observed ratings in the left panel of figure 8.7.

8.4.3 Continuous Model

For the continuous model, I assume that readers perceive the degree to which a test letter resembles each alternative. The percept of a stimulus

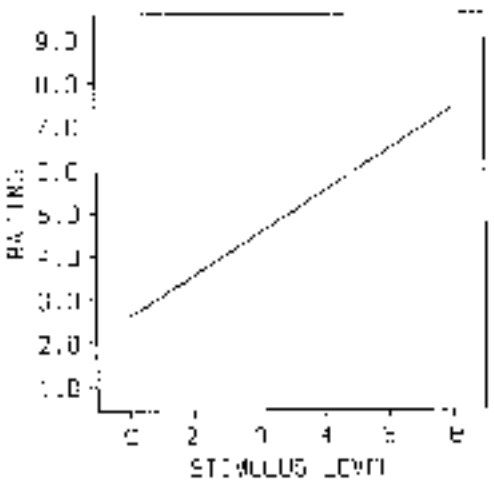

Figure 8.9
Hypothetical mean rating responses predicted by both the discrete and continuous models as derived from figures 8.8 and 8.10.

toward the *c* end of the continuum will be more *c*-like than that of a neighboring stimulus toward the *e* side of the continuum. Although there is continuous information available, we can also expect a distribution of rating judgments for each of the test stimuli because of random variability in the perceptual, memory, or response systems. As illustrated in figure 8.10, the continuous model predicts a systematic and continuous change in the distribution of percepts and the corresponding ratings across the six test letters. The continuous model also predicts that the mean of the rating responses (marked by arrows in figure 8.10) will change continuously with changes along the stimulus continuum. These mean ratings in figure 8.10 also fall on the line in figure 8.9, and are consistent with the observed results shown in the left panel of figure 8.7.

As we have now seen, both the discrete and continuous perception models can predict the observed means of the rating responses shown in figure 8.7. Thus mean rating judgments alone are not capable of distinguishing between the two models. This situation would be described as a lack of identifiability. That is, the models cannot be distinguished based on the data in figure 8.7. The models, however, can be distinguished on the basis of the distributions of rating responses, which are predicted to differ. Figures 8.8 and 8.10 illustrate the overall form of the predictions. As can be seen, although the means of the rating responses (marked by the arrows) can be identical for the two models, the distribution of rating responses will necessarily differ. For example, for stimulus level 3, the

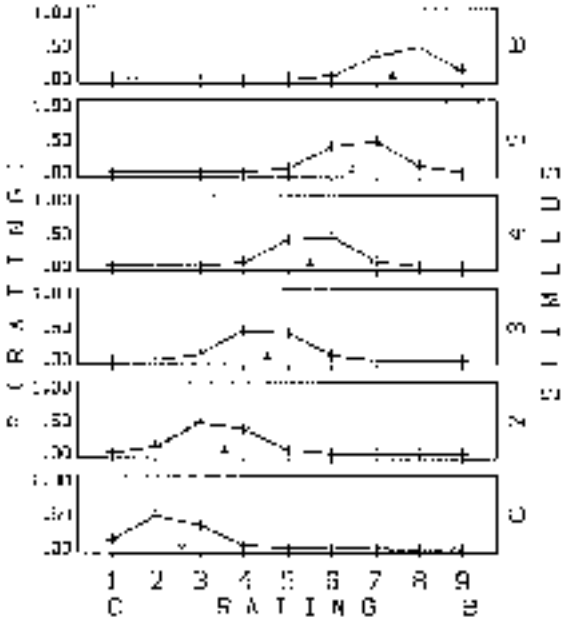

Figure 8.10
According to the continuous model, the rating responses can be described by just a single distribution for each of the six stimulus levels along the stimulus continuum. For each stimulus, rating responses are based on the response probabilities shown for that stimulus. The downward pointing arrows mark the mean response rating for each of the six stimuli.

discrete model predicts a two peaked (bimodal) distribution with a central trough (figure 8.8), whereas the continuous model predicts a distribution with a single peak (figure 8.10).

8.4.4 Experimental Test

Returning to the initial letter-rating task, we can analyze the observed distributions of the rating responses. The subject provided seventy ratings for each of the six test letters. The circles in figure 8.11 show the rating judgments given by the subject. As can be seen, the distributions of the observed ratings are unambiguously characterized by having a single peak as predicted by the continuous model. Consider the third stimulus level on the *c-e* continuum. Figure 8.11 shows that the subject consistently rated this stimulus as a four along the continuum between *c* and *e*. As shown in figure 8.8, the discrete mode would predict that the subject should have rated it some of the time toward the *c* end and some of the time toward the *e* end of the rating continuum.

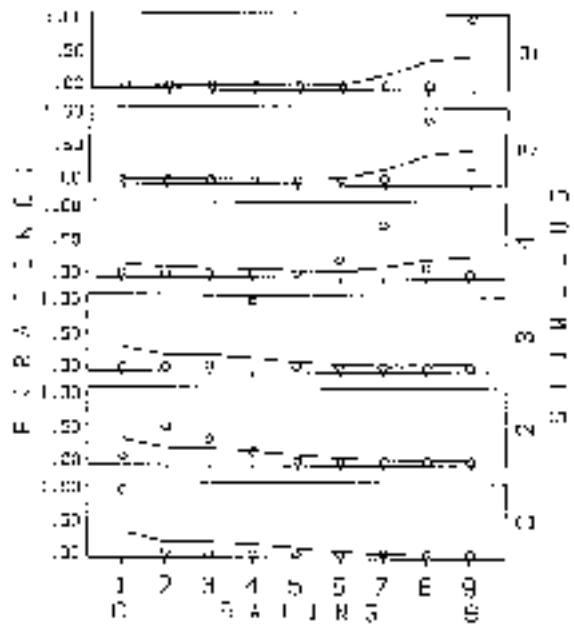

Figure 8.11
Observed (dots) and predicted (lines) proportions of the nine possible rating responses for each stimulus level. Predictions of the discrete feature model (DFM).

In order to determine whether the observed distributions of the ratings were best fit by the continuous or the discrete model, the two models were formulated to predict the distributions of ratings. That is, instead of using the hypothetical distributions shown in figures 8.8 and 8.10, I found the distributions for each model that would do the best job of predicting the observed rating judgments. This permits a quantitative comparison between them. Figure 8.11 also shows the predictions (lines) of the discrete perception model. Figure 8.12 shows the same data along with the predictions of the continuous perception model. The continuous model does a much better job of fitting the observed distributions of ratings. In terms of the differences between the observed and predicted rating distributions, the description given by the continuous model was about fifteen times better than that given by the discrete model. (Later in this chapter, I will discuss methods for fitting models to observed data and evaluating the fits.)

Once the predictions shown in figures 8.11 and 8.12 were obtained, the predicted mean rating response for each stimulus could be calculated. For example, for the 6 stimulus, the line in the top panel of figure 8.11 predicts that the rating should be 9 on about 30 percent of the trials, 8 on 15 percent, and 7 on 15 percent. The mean rating for this stimulus, based on this predicted distribution of rating responses is 8.35. The predicted mean ratings for each stimulus worked out in this way for both figures 8.11 and 8.12 are shown as lines in figure 8.13. The gulf between the two models is clear. The continuous model predicts the observed mean ratings accurately, whereas the discrete model does not. Given these tests of the two models, we can conclude that letter features are perceived continuously rather than discretely. What is important for our purposes is that this conclusion required computational modelling to provide a definitive test of a long-standing assumption about the processing of letter features.

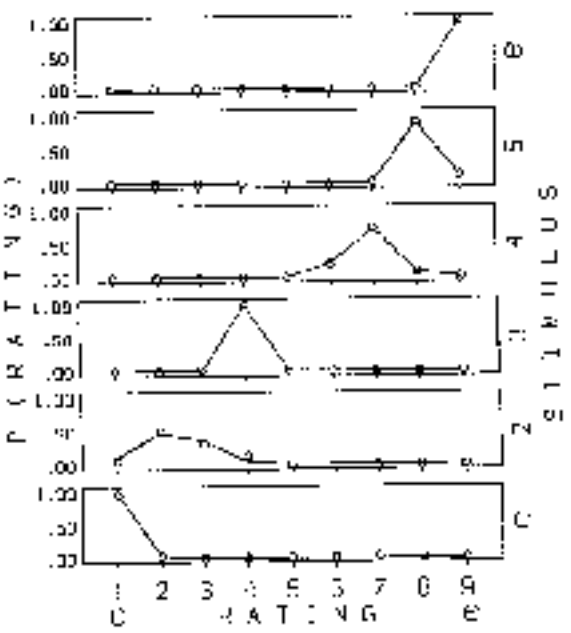

Figure 8.12
Observed (points) and predicted (lines) proportions of the nine possible rating responses for each stimulus level. Predictions of the continuous feature model (CFM).

8.3 Multifactor Experiments

To this point, our experimental and theoretical investigations have been relatively simple compared to reality. In the last study I varied just a single feature although I believe that letters are characterized by multiple features. And although I described this simple experiment as a test of whether features are perceived as discrete or continuous, the results are consistent

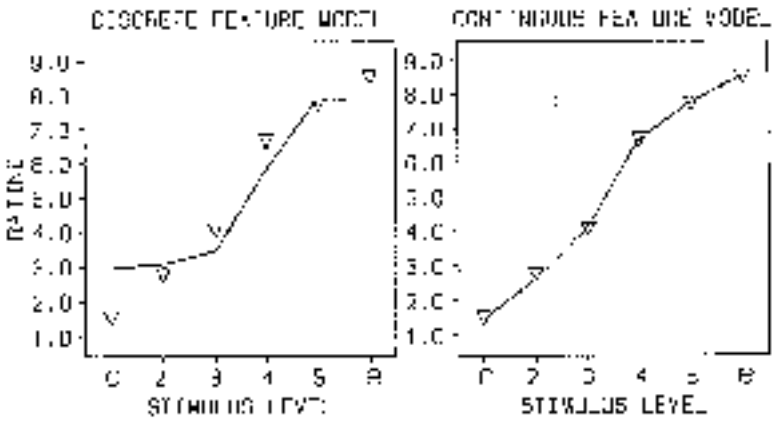

Figure 8.13
Observed (triangles) and predicted (lines) meaningfulness judgments as a function of stimulus level. Left panel shows the predictions of the discrete model. Right panel shows the predictions of the continuous model.

with both feature and template theories. That is, the results tell us only that subjects responded to the information in the task as if it were continuous.

If we design our experiments to be somewhat more realistic, we should systematically vary two or more features independently to try to isolate the influence of each feature. Furthermore, following the logic used earlier, we should also vary the degree to which each feature is present. To illustrate this type of experiment, I will describe and analyze a more complex pattern categorization task in which I vary several sources of information.

Two categories, *G* and *Q*, were chosen as the alternatives in a letter-processing task. Let us evaluate the uppercase letters Q and G from a featural perspective. The letter Q has a closed oval, while the letter G has an open one. The letter Q has an oblique line, while the letter G has a horizontal one. We can create novel test letters by independently modifying these two features and varying the degree to which each feature resembles its normal appearance in the letters.

For the oval, seven levels of gap size were created 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 angle of the line crossing the oval was varied between the horizontal and 11, 21, 29, 36, 51, and 61 degrees measured from the horizontal. In this experiment, the two variables of gap size and line angle are called "factors," and each of these factors has seven possible values or "levels." In a factorial experimental design with two factors, each level of one factor is combined with every level of the other factor. The resulting 49 test letters are shown in figure 8.14, where the 7 columns correspond to the levels of the gap size factor, and the 7 rows correspond to the line

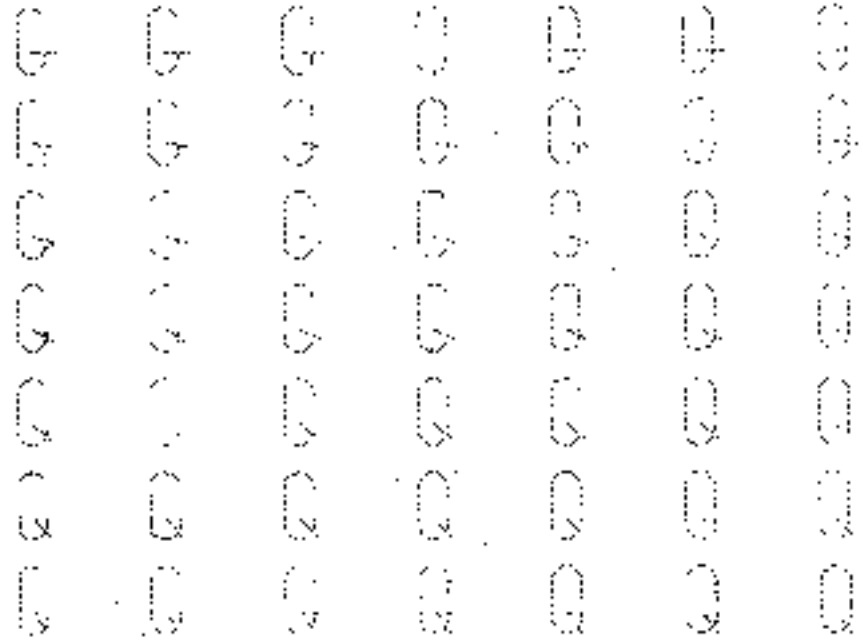

Figure 8.14
Forty-nine Q/G test letters created by varying the degree to which the the straight line is oblique (row variable) and the degree to which the oval has a gap (column variable).

angle factor. These stimuli make up the factorial design in this experiment. As in the single-factor test study, we can ask a subject for either categorization or rating responses.

Massaro and Hary (1986) carried out this experiment. In the categorization task, the 49 test letters were each presented 12 times for 400 milliseconds in random order to each of 9 subjects. Subjects categorized each test letter as *G* or *Q*. The probability (proportion) of *Q* responses, *P*(*Q*), for each test letter was the dependent variable. Given that the *P*(*Q*) and *P*(*G*) proportions must sum to one, a measurement of *P*(*Q*) for each test letter completely represents the categorization responses. Thus the results consist of 49 independent proportions, *P*(*Q*) for each subject.

8.6 Models of Recognition

8.6.1 Template Model

A template account of this task appears to have very little predictive power. Each of the forty-nine test stimuli is a new holistic event and

recognition cannot be predicted on the basis of the two separate properties, gap size and line angle. What is important for recognition is the overall global degree of match of a test letter with a subject's ideal template representations of Q and G. It could be assumed that the upper left and lower right letters in figure 8.14 correspond to a subject's templates and that recognition of each of the test letters is related to the degree of match between the test letter and the two templates. In this scheme, a perceiver must compute the amount of overlap between a test letter and each template and categorize the letter as the alternative with the greatest overlap. Thus every letter in figure 8.14 could be categorized as a G or a Q based on whether it was a better match with a G or a Q template.

This version of the template theory predicts a consistent division of the forty-nine letters in figure 8.14 into two sets. But this prediction conflicts with the variability of human nature. My daughter's softball team provided a recent example. A batter would miss several pitches by a mile and then get a perfect hit, or the pitcher would throw several strikes and then miss the plate entirely, and so on. Similarly, in pattern recognition, we can expect that a stimulus will not always receive the same categorization response. One place where variability might occur in a template model is in the decision operation. For example, if decisions are based on a relative goodness-of-match rule (RGR), then the probability of a Q response is equal to the goodness of match of the test letter with the Q template relative to the sum of the goodness-of-match values of the test letter with the Q template and with the G template. This rule, in equation form, is

$$P(Q) = \frac{M(Q)}{M(Q) + M(G)} \tag{8.1}$$

where $M(a)$ is the goodness of match of a test item with template a.

One justification of the RGR is the probability matching that is often observed in decision making. If people are asked to predict which of two events will occur after having experienced one event 70 percent of the time and the other 30 percent, they do not always choose the most frequent event. They tend to choose the more frequent event about 70 percent of the time, that is, they probability match. The RGR represents a similar idea where the probability of a response is based on the relative goodness of match.

The RGR also captures what appears to be a reasonable relationship between stimuli and responses in pattern recognition. Consider a case in which a test letter gives a reasonably good match only to the alternative Q, and a second case in which the test letter gives about the same match to Q but also a somewhat less good match to the letter G. If the absolute match were the important value for decision, the probability of classifying

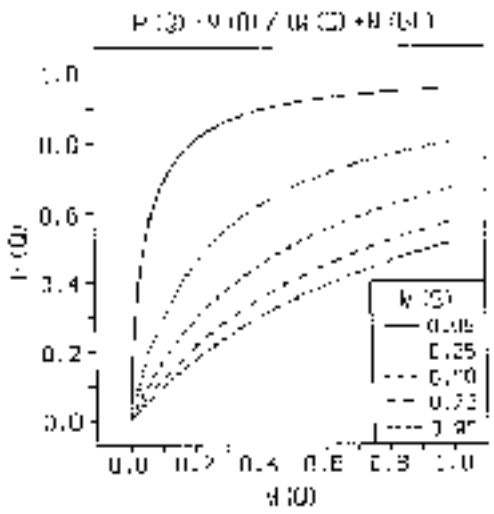

Figure 8.15
Predicted P(Q) values as a function of some hypothetical values of $M(Q)$ and $M(G)$.

the letter as Q should be 1 in both cases. However, the RGR predicts that $P(Q)$ should be much less in the second case, and the test letter will be categorized as G some of the time. This idea is illustrated in figure 8.15, which gives some predicted $P(Q)$ values as a function of some values of $M(Q)$ and $M(G)$. As can be seen in the figure, $P(Q)$ depends on both $M(Q)$ and $M(G)$. If $M(G)$ is small (e.g., 0.05), then $P(Q)$ increases very quickly to near 1 with increases in $M(Q)$. On the other hand, if $M(G)$ is large, then $P(Q)$ increases much more slowly to about .5 with increases in $M(Q)$. More generally, $P(Q)$ becomes more likely to the extent that $M(Q)$ exceeds $M(G)$.

I have ked out some predictions of the template model by using the RGR to allow probabilistic categorization of the same stimulus. Another aspect of these predictions is not realistic, however. Given that the response probability is still based on the objective overlap between the stimulus and template, the model must make the same prediction for each subject. But figure 8.16 indicates that subjects differ from one another. Upon reflection, we can see that each of us has had unique life experiences, and there is no reason to believe that two individuals will behave identically in the same situation. More specifically, why should my ideal letter Q be the same as yours? Given different ideal letters, the way in which people categorize the stimuli in figure 8.14 as Qs and Gs should also differ. Thus any reasonable template model cannot be based on objective overlap that can be measured directly in the stimuli it must

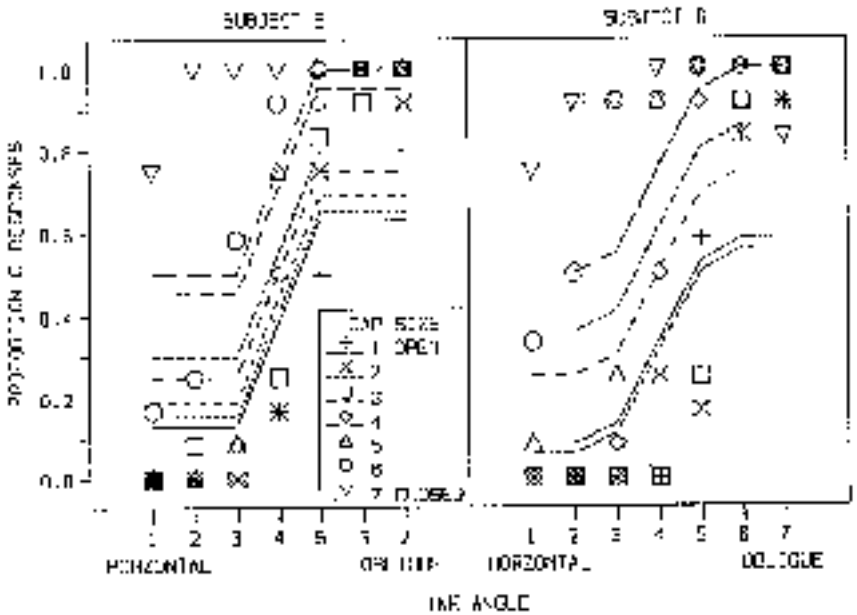

Figure 8.16
Observed (points and predicted (lines) proportion of Q responses for 49 Q/O test letters as a function of the line angle and gap size. (Predictions of the discrete feature model, DFM, for the results of two subjects; Massaro and Hary 1986)

assume subjective overlap that will be unique for each subject. In this case, each test letter requires independent estimates of the amount of subjective overlap with the templates for each subject. Thus the template model can only predict the categorization probability for a test letter by estimating a subjective overlap value for that test letter. This puts us in the unhappy situation in which we cannot predict performance on the test letters without first measuring how subjects respond to each of the 49 stimuli. With this stipulation, we are left without a testable template model for the G-Q task.

The lack of testability of the template model reveals a fundamental incompatibility between theory and experiment. Theories depend on empirical tests. A theory is not worthwhile unless it is testable.

8.6.2 Discrete Feature Model

Although our experiment cannot test template models, it is ideal for testing feature models. I will describe a discrete (binary) feature model (DFM) for perceiving Q versus O, based on gap size and line angle features. But before doing so, you might wonder why I make the effort to test such a

model when the idea of discrete features was already falsified in the G-Q study. There are several justifications. First, one goal of science is generalization across a variety of specific instances. It is important to know if conclusions reached in the single-factor G-Q study can be extended to a two-factor G-Q study. Second, a combining or integration process might occur in the two-factor study that is not needed in the single-factor G-Q study, and the addition of an integration process to combine information about two or more features might limit the perceiver to only discrete information about each feature. Third, we will see that the more complex experimental situation allows us to derive a more complex discrete feature model that warrants empirical test because the more complex model might give accurate predictions even though a simpler version has already been falsified.

With just a single feature, we might postulate two psychological processes: evaluation and decision. To illustrate, in the G-Q recognition task described above, the horizontal line feature must be evaluated and then a categorization or rating decision must be made. But given multiple features, such as in the G-Q recognition task just described, there is the possibility that an additional integration operation occurs after evaluation but before decision. This integration operation combines or integrates the several features or sources of information that have been evaluated before a decision is made. This 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 gap size and line angle components of the test letter. *Integration* is defined as a combination of the information made available by the evaluation process. Decision converts the outcome of integration into a response.

The three processes are illustrated in figure 8.17. I illustrate the processes as overlapping because, although they are necessarily successive, one process could begin before a previous process is finished. (For example, the evaluation of some features might continue even while the integration process has started for others.) To develop a model, each of these processes must be exactly specified in order to make predictions of performance.

8.6.2.1 Elaborating the Presumed Operations

Memory is an essential component of pattern recognition because the current stimulus pattern has to be compared to the recognizer's memory of previous patterns. One possible type of pattern memory is a set of summary descriptions of the meaningful patterns. These summary descriptions are called "prototypes" or categories, and each prototype is a description

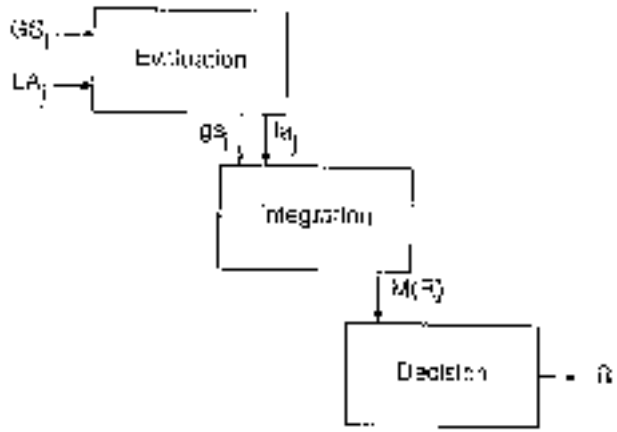

Figure 8.17
Schematic representation of the three stages involved in perceptual recognition. The three stages are shown to proceed left to right, in time to illustrate, but is necessarily successive but overlapping processing. The sources of information are represented by uppercase letters GS_i and LA_i. The evaluation process transforms these sources of information into psychological values corresponding to the features of the prototype indicated by lowercase letters gs_i and la_i. These sources are then integrated to give $M(Q)$ and $M(G)$. The evaluation and integration processes occur for each prototype, with the decision process getting $M(Q)$ and $M(G)$ values from all of the prototypes. The decision operation maps these values into some response, R, such as a rating or a letter identification response.

of the ideal feature values that a pattern should have if it is a member of that category. The exact form such prototype descriptions might take in the mind is not known and may never be known. However, the representation in memory must be compatible with the representation that results from the transduction of the stimulus. Compatibility is necessary because the perceiver must compare the information provided by the stimulus to some memory of the prototypes or categories.

The first operation in figure 8.17 evaluates the stimulus to determine the values of the features, that is, which features are present and which are absent. In developing the discrete feature model (DFM) for the c-e experiment (section 8.4.2), I assumed that a given stimulus is not always perceived in the same way. For the Q-G stimuli, this means that for a particular stimulus in column i and row j of figure 8.14, a subject might see the gap as closed with probability p_i and as open with probability $1 - p_i$. Similarly, the line angle might be seen as horizontal with probability q_j and as not horizontal with probability $1 - q_j$.

The second operation, feature integration, compares each of the prototypes to the stimulus features, and the outcome is a measure of how much each prototype matches the stimulus. In order to integrate information about the degree of match of each feature into an overall goodness of match for a prototype, the various features must be evaluated using the same measurement scale. In the computational model's that I develop below, each feature is given a numerical measure of its degree of match that serves this measurement function, and for each prototype, the sum of these measurements represents how well the prototype matches.

The third operation, decision, must select a response based on the outcome of integration. The decision operation might simply select the prototype or category with the best goodness of match, as I suggested for the template model (section 8.6.1), and I will discuss this possible decision rule later. However, for now I will again use the relative goodness-of-match rule that I proposed for the template model (section 8.6.1, equation 8.1). With this rule, the decision in this DFM is probabilistic, based on a prototype's relative goodness of match, where the relative goodness of match of a prototype is the prototype's overall goodness of match with the stimulus, divided by the sum of the goodness of match information for all the prototypes. In Pandemonium terms, we might say that it is not how loudly some demon is shouting but rather the relative loudness of that demon in the crowd of demons. This relative goodness of match gives the probability that the stimulus is identified as an instance of a particular prototype.

The three operations between presentation of a pattern and its categorization, as illustrated in figure 8.17, can be formalized mathematically. Assume that the hypothetical Q and G prototypes are defined by the features shown below. Let us also assume that the output of the feature evaluation process for each prototype is 1 (indicating that the feature is present) or 0 (indicating that the feature is absent). For example, the third level of gap size shown in figure 8.14 might be perceived as closed, which would match the description of the Q but not the G prototype.

Q: Closed gap and not horizontal angle
G: Not closed gap and horizontal angle

Feature integration, as shown in figure 8.17, consists of counting the number of matching features for each prototype. Finally, assume that in the decision operation, the count of the number of matching features for each prototype is evaluated relative to the sum of the match counts for all prototypes. This relative goodness of match gives the probability that the stimulus is identified as an instance of a prototype.

Limiting the prototypes to only two features, gap size and line angle, in no way implies that I believe that these are all the features of these two letters. But, given that these two are the only features being manipulated in the experiment, the prototypes can be represented in simplified form with just these features.

To illustrate these ideas, consider the fourth level of gap size and the fourth level of line angle in figure 8.14 (column 4, row 4). Feature evaluation might give a 1 (feature present) for the closed gap feature and a 1 for horizontal angle feature. In this case, the feature integration sum for these two features would be 1 + 0 for the Q prototype (match with the closed gap feature value, but mismatch with the not horizontal angle feature value) and 0 + 1 for the G prototype. If $M(Q)$ and $M(G)$ correspond to the number of matches to the Q and G prototypes, respectively, then both $M(Q)$ and $M(G)$ would be equal to 1. If Q and G are the only valid response alternatives, the decision operation determines their relative goodness as in equation 8.2, where $P(Q)$ is the predicted probability of a Q response to a particular test letter shown in figure 8.14. This relationship means that my DFM and template models assume the same decision operation. They differ in terms of evaluation and integration operations. For this example, $P(Q) = 0.5$.

$$P(Q) = \frac{M(Q)}{M(Q) - M(G)} = \frac{M(Q)}{2} \tag{8.2}$$

In general, $M(Q)$ can be 0, 1, or 2, depending on whether the test stimulus is evaluated as having 0, 1, or both of the Q features. Because Q and G are defined as differing on both of the binary features, $M(G) = 2 - M(Q)$. Substituting $2 - M(Q)$ for $M(G)$ in equation 8.2 leads to the simplified form of equation 8.2, $P(Q) = M(Q)/2$, as shown above.

To derive the predictions of the DFM given by equation 8.2, it is necessary to determine the likelihood of obtaining the different $M(Q)$ values for a test stimulus. Remember that in the DFM a feature either matches a prototype or not; there is no meaningful territory in between. However, in the evaluation of the stimulus, there is some probability that a feature will be assigned each of its two alternative feature values, for example, closed or not closed for the gap feature. Also, the probability of assigning each of the values of a feature can be assumed to vary with the characteristics of the test letter. Suppose that p_i is the probability that the gap size feature is classified as closed, given the ith level of gap size in figure 8.14. Similarly, suppose that q_j is the probability that the line angle feature is classified as not horizontal, given the jth level of line angle. For this stimulus, table 8.1 gives the probabilities of the four possible feature value assignments that can occur for a particular stimulus as a result of this evaluation variability. Note that $M(Q) = 2$ for the upper left entry of table 8.1, $M(Q) = 0$ for the lower right entry, and $M(Q) = 1$ for the remaining two entries. This means that when the same stimulus is presented several times, the $M(Q)$ value can vary. And because $P(Q)$ depends on $M(Q)$, we have to compute the overall $P(Q)$ probability by first computing $P(Q)$ using equation 8.2 for each of the possible outcomes in table

Table 8.1
The probabilities of the four possible outcomes of the feature evaluation process for the decision feature model (DFM)

	Line Angle	
Gap size	Not horizontal	Horizontal
Closed	$p_i q_j$	$p_i(1 - q_j)$
Not closed	$(1 - p_i)q_j$	$(1 - p_i)(1 - q_j)$

8.1, and then computing an average of these $P(Q)$'s based on the probability of each of them as shown in table 8.1.

Suppose that we know the values of p_i and q_j. Then these probabilities can be used to predict the overall $P(Q)$ for a given test stimulus, GS_i, LA_j with gap size GS_i and line angle LA_j, as follows:

$$P(Q|GS_i, LA_j) = (1)p_i q_j + (.5)p_i(1 - q_j) + (.5)(1 - p_i)q_j + (0)(1 - p_i)(1 - q_j) \tag{8.3}$$

where $P(Q|GS_i, LA_j)$ is the probability of a "Q" response, given stimulus GS_i, LA_j, and where i and j index the levels of the gap size and line angle features, respectively. Each of the four terms in equation 8.3 represents the probability of one of the four possible feature value outcomes multiplied by $P(Q)$, the probability of a Q identification response, given that outcome. For example, the first term in equation 8.3 ($(1)p_i q_j$) corresponds to the outcome in the top left entry in table 8.1 in which both features are classified as matching the prototype for Q (and not matching the prototype for G). In this case, $M(Q)$ is 2, and $P(Q)$ is 1. Because this outcome occurs with probability $p_i \times q_j$, the probability of a Q response is, therefore, $1 \times p_i \times q_j$. The other three terms in equation 8.3 correspond to the other three entries in table 8.1 and can be described similarly.

Using algebraic rules, equation 8.3 reduces to:

$$P(Q|GS_i, LA_j) = \frac{p_i + q_j}{2}. \tag{8.4}$$

The form of equation 8.4 shows that the probability of a Q response is equal to the average of the probabilities of evaluating the gap size feature as closed and the line angle feature as not horizontal. The value of $P(Q)$ can vary between 0 and 1, as it should. When p_i and q_j are zero, $P(Q)$ is zero; when p_i and q_j are both one, $P(Q)$ is one; and if each feature is detected about half the time, $P(Q) = .5$. Finally, because there are only two response alternatives,

$$P(G|GS_i, LA_j) = 1 - P(Q|GS_i, LA_j). \tag{8.5}$$

The predicted response probability in equation 8.4 is based on p_i and q_j, but we do not know the ideal values of these variables, and we do not know how accurate the model is until we can compare its predictions with the observed results. Before proceeding with tests of the model, it is therefore necessary to describe how values for these variables are determined. This involves the concept of free parameters and their estimation.

8.6.2.2 Free Parameters and Their Estimation

In terms of the DFM, we do not know the p_i and q_j values because we cannot state ahead of time how changes in the letter features will change these probability values. However, the DFM makes some strong predictions that allow it to be quantitatively tested. I will assume that in the DFM model, the probability of classifying a particular feature depends on the physical characteristics of that feature, but not on the other features. Without this *independence* assumption, the DFM would be essentially untestable, in the same way a template model is untestable. With the independence assumption, we see that there can be a unique value of p_i for each of the seven levels of gap size that is unaffected by line angle. Similarly, there can be a unique value of q_j for each level of line angle. We do not know what these values are, and we must use the results given by a subject to find them. This process is called "parameter estimation."

Most computational or quantitative models have a set of free parameters. A free parameter in a model is a variable that cannot be exactly predicted in advance. The actual performance of a subject is used to set the value of this variable. Predictions of behavior cannot be exact or even very accurate without first knowing something about what is being predicted. Taking our ambiguous letters in the *c-e* experiment as examples, we cannot know exactly how often a given person will categorize one of the letters as a particular alternative. Nor could we know in advance the exact rating a subject would give one of the letters. Notice that the two subjects in figure 8.16 give similar but not identical results. We can know that one letter might be more likely to be identified as a *G* than another, but we do not know how much more. This uncertainty would preclude the quantitative test of letter recognition models if we were not able to estimate free parameters.

In parameter estimation, we use our observations of the subject's behavior to estimate the values of the free parameters of the model being tested. Because we want to give every model its best shot, our goal should be to find the values of the parameters that maximize how accurately the model is able to account for the results. This is called "maximizing the goodness of fit" of the model. When we compare competing models, each model should be predicting as well as it can to increase the fairness of the test.

The DFM has a $P(Q)$ prediction (equation 8.4) for each of the 49 conditions in the *Q-G* experiment. The 49 equations have seven different values of p_i because there are 7 levels of gap size, and analogously for q_j and the 7 levels of line angle. Thus we have 49 equations with 14 free parameters, the 7 p_i and 7 q_j values. Finding values for these 14 free parameters allows us to predict the 49 observations. Again, we want the values that maximize the fit of the predictions to the observations. Finding the optimal parameter values can be done by computer using a parameter search algorithm. The logic of this estimation procedure is to try out a variety of values to find those that minimize the differences between the predicted and observed values. Estimating the 7 p_i and 7 q_j values in equation 8.4 will give the predicted values that we need to compute the 49 values of $P(Q)$. This search is made easier by giving a permissible range of values. In the DFM, for example, the p_i and q_j values are probabilities and thus must be between 0 and 1.

A criterion that is often used to measure the goodness of fit is the *root mean square deviation* (RMSD) between the predicted and observed values. The best fit is that which gives the minimal RMSD. The RMSD is computed by (a) squaring the difference (or deviation) between each predicted and observed value (this makes all differences positive and also magnifies large deviations compared to small ones); (b) summing the squared deviations across all conditions (49 in the *Q-G* task); (c) taking the mean (dividing the sum by 49 in the *Q-G* task); and (d) taking the square root of this mean. This RMSD can be thought of as a standard deviation of the differences between the 49 predicted and observed values. (Wickens, chap. 12, this volume, for a discussion of the standard deviation.) Thinking of the RMSD as a standard deviation will also help us compute a benchmark that can serve as a standard for evaluating how well a particular model does. All of the differences would be zero in a completely accurate model, and the RMSD would be zero. The RMSD increases as the differences between observed and predicted values increase. In general, the smaller the RMSD value, the better the fit of the model.

We are impressed with a model to the extent that its predictions go substantially beyond the data and the assumptions used to construct the model. Consider a hypothetical model that predicts that the probability of an *e* response in the *c-e* experiment is a linear function of the length of the horizontal line of the test letter. This means that if we graph the probability of an *e* response, $P(e)$, on the *y*-axis against the length of the horizontal line on the *x*-axis, we predict that the observed data points for $P(e)$ should fall along a straight line. Note also that because this model predicts a straight-line relationship, the model has two parameters, a slope and an intercept. In comparing the model to data, it may not be possible to predict the slope and intercept parameters in advance. That is, the model may predict only that $P(e)$ should increase linearly as the length of the

horizontal line increases but may not predict the actual values of the slope and intercept (though it does predict that the slope should be positive). If we carry out an experiment using stimuli with just two horizontal line lengths, we will have only two data points, and the model will necessarily fit the data perfectly. A straight line can always be fit perfect by through two points. In this example, the slope and intercept are free parameters because they are free to take on almost any positive values; they have not been specified or predicted by the model. And we would need three or more data points before we could evaluate this model.

We become more impressed with a model to the extent that the number of points being predicted exceeds the number of free parameters being estimated. I have to caution you, however, that there is no convenient measure of how well a model describes a set of outcomes relative to the number of free parameters needed to predict those outcomes. For example, if we have two models, one with two free parameters and a second with three free parameters, the model with three parameters will usually do better in matching the results. But this is only because, in a sense, the three-parameter model is less specific and more open ended in its predictions, and it can be adjusted more easily to any observed set of data. For example, if in the Q-G experiment we have data points for three or more stimuli, then relative to a single-line model, we can expect that a model that predicts that the data points should fall along two connected line segments, each with its own slope parameter, would give a better match to the data. We have to exercise extreme caution when the model that fits the data best has the most free parameters. In testing among models, we are usually satisfied when one model does better than other contenders (e.g., has the smallest RMSD) and also has the same number or fewer free parameters.

Returning to the DFM and the results of the Q-G experiment, with seven levels of each of two features, we need a q_i parameter for each level of line angle, and a p_i parameter for each level of gap size in order to predict the 49 data points. I estimated the 14 free parameters to minimize the RMSD between the 49 observed and predicted data points. The model fit was carried out separately on each subject's results. Fitting the DFM to each individual permits the model to describe each subject while allowing for differences between subjects, insofar as these can be captured by the differences among the 14 parameters.

The predictions of the DFM model are given in figure 8.16 for two subjects. The data points are the observations, and the lines correspond to the model predictions, where each line represents predictions for one level of gap size. Figure 8.16 shows the parameter values for the two subjects that maximized the goodness of fit of the DFM. The predictions do not capture the trends in the data very well. The RMSDs for the two subjects were 0.210 and 0.224, respectively. These large RMSDs are not usually

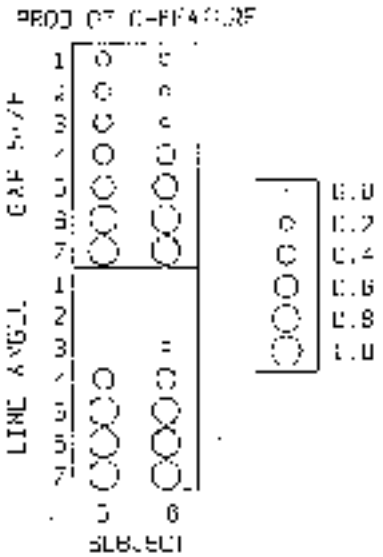

Figure 8.16
Fit of the DFM parameter values to the probabilities of evaluating the gap size is not closed and that line angle, for the seven gap sizes and seven line angles for the two subjects shown in figure 8.18. The scale on the right shows the relationship between area and parameter value.

considered a good fit when the data are proportions. Figure 8.16 also reveals that the failure of the DFM is systematic. The predictions are a set of parallel lines, whereas the observations, particularly the data for the open and closed oval stimuli, resemble an American football—wide in the middle and narrowing at the ends. The parallel lines prediction means that the influence of a change of the line angle feature on P(Q) is constant across all levels of the gap size feature. But in contrast with the DFM model predictions, the observations show that a feature's influence increases as information about the other feature becomes more ambiguous.

Feature models predict more than they assume. That is, the DFM made predictions for 49 data points based on estimates of 14 free parameters. In this case, the feature model is more parsimonious than the template model (which required a free parameter for each of the 49 stimuli). However, because the DFM predictions are not very accurate, and because it is also important to test other models against the same results, I turn to a model that assumes readers have continuous featural information available.

8.6.3 Fuzzy Logical Model of Perception

Like the general model in figure 8.17, the fuzzy logical model of perception (FLMP) assumes feature evaluation, feature integration, and decision

(Oden and Massaro 1978). 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 prototype descriptions. Thus the FLMP differs from the DFM in terms of having continuous rather than discrete information about the features. For the common metric that is necessary to integrate information about several features, truth values from fuzzy logic (Zadeh 1965) are used because they provide a natural representation of the degree of match. Fuzzy truth values lie between zero and one, corresponding to a hypothesis being completely false and completely true. The value 0.5 corresponds to a completely ambiguous situation (e.g., a line that is halfway between horizontal and not horizontal), whereas 0.7 would be more true than false, and so on. Fuzzy truth values, therefore, not only can represent different kinds of information; they can represent continuous rather than just discrete information.

In terms of the three operations represented in figure 8.17, fuzzy truth values arise in the first operation, feature evaluation, and represent the degree to which particular characteristics of a test stimulus support the presence of particular features. Also, feature evaluation is deterministic, that is, the evaluation process always gives the same truth value for a particular stimulus. This is in contrast to the DFM, where a feature value was assigned with some probability. Suppose we represent the gap size and line angle features of the stimulus by the symbols GS_i and LA_j. In feature evaluation, the physical input is transformed to a psychological value, and is represented in lowercase, for example, GS_i would be transformed to gs_i, the degree to which the gap size supports the gap size feature of Q. With just two alternatives, Q and G, we can make the simplifying assumption that the degree to which the gap size supports the alternative G is $1 - gs_i$. Feature evaluation would occur analogously for the line angle feature, LA_j. Feature integration consists multiplying the feature values supporting a given alternative. Accordingly, a second difference between the FLMP and DFM is multiplicative, as opposed to additive, integration. If gs_i and la_j are the values supporting alternative Q, then the total support for the alternative Q, $M(Q)$, would be given by the product of gs_i and la_j,

$$M(Q) = gs_i la_j \tag{8.6}$$

The third operation is decision, which gives the relative degree of support for each of the response alternatives. In this case, the probability of a Q response given stimulus GS_i, LA_j is equal to the total support for Q divided by the sum of the support values for all relevant alternatives (e.g., Q and G), or, in the DFM model equation

$$P(Q|GS_i, LA_j) = \frac{M(Q)}{M(Q) + M(G)} = \frac{gs_i la_j}{gs_i la_j + (1 - gs_i)(1 - la_j)} \tag{8.7}$$

To summarize, the FLMP and the DFM differ with respect to two of the three processes in figure 8.17: the evaluation process produces deterministic and continuous values in the FLMP, versus probabilistic binary feature values in the DFM; and the integration process is multiplicative for the FLMP, versus additive for the DFM. Both models use the same decision rule. In both cases, the FLMP preserves processes that are more efficient or optimal than those assumed by the DFM. It is obvious that having continuous information is an advantage over being limited to discrete information. Knowing how much an item costs is more informative than simply knowing the item is expensive. Although not as intuitive, it is easily proven that multiplying two sources of information in the FLMP is more optimal than simply adding the sources in the DFM (Massaro 1987; see section 8.10.3).

As in the DFM, 14 free parameters are necessary to fit the FLMP to the 49 data points: 7 parameters for each level of gap size and line angle. The parameters represent the degree to which the gap size and line angle features match the feature values of the Q prototype. Because the two models differ in their structure, these parameter estimates for the FLMP will not be the same as for the DFM. The predictions of the model (for the same two subjects fit by the DFM in figure 8.16) are given in figure 8.19. The predictions do very well in capturing the trends in the data. The RMSDs for the two subjects were 0.0412 and 0.0543, respectively, about seven and four times better than the DFM predictions for these two subjects.

This analysis reveals that letter recognition can be described in terms of the evaluation and multiplicative integration of continuous independent features with respect to prototypes in memory, and that a decision is based on the relative goodness of match. As noted earlier, the FLMP and the DFM differ in two assumptions: continuous versus discrete feature information, and multiplicative versus additive integration. Other research has demonstrated that both differences are important (Massaro 1987). A multiplicative combination of discrete features or an additive combination of continuous features alone cannot predict the results.

Figure 8.20 illustrates the parameter values for the two subjects that maximized the goodness of fit of the FLMP. The values show that both independent variables influenced the judgments in the expected manner. Decreasing the gap size in the oval and making the line angle more oblique increased the support for the alternative Q. Some of the values appear to change nonlinearly from one level of a feature to the next level. Observing the growth curve to the values only two values

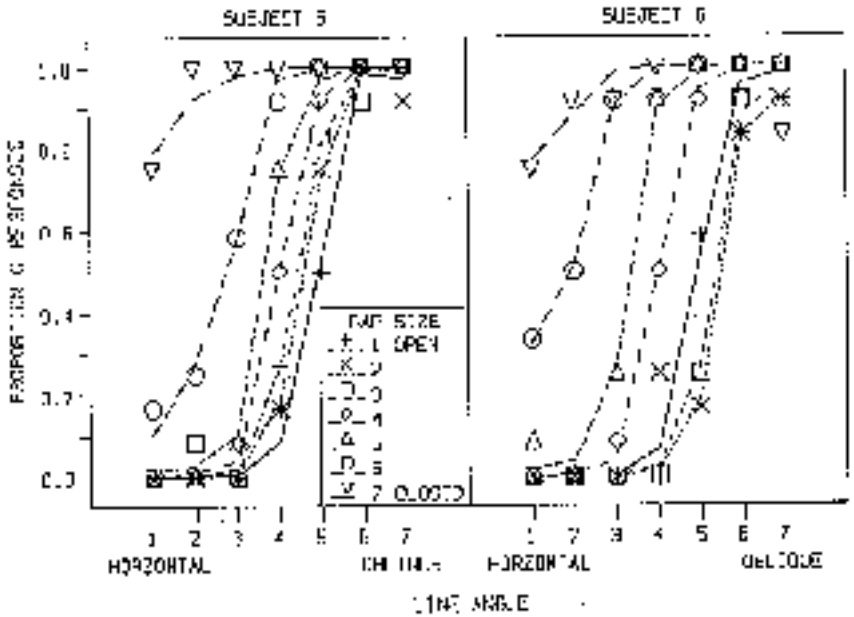

Figure 8.19
Observed (points) and predicted (lines) proportion of Q responses for 49 Q/G test letters as a function of the line angle and gap size. (Predictions of the fuzzy logical model of perception, FLMP, for the results of two subjects from Massaro and Hary 1986.)

for the seven levels of a feature continuum might describe the results just as well. This is not the case because the average RMSD increases to .342 for this model.

8.6.3.1 Benchmark Measures of Goodness of Fit

Even if a model is a correct description of the psychological processes involved, we cannot expect it to fit observed results perfectly. As mentioned earlier, models must be probabilistic or have built-in variability to be consistent with the variability that people display. One way that models can do this is the way in which the DFM and FLMP do—by specifying response probabilities. Then, if one simulates the behavior of a model several times, the observed response proportions will vary simply because of sampling variability, just as the observed number of heads will vary if several coins are tossed repeatedly. For example, suppose a model predicts that $P(Q) = 0.6$. Given this probability, we cannot expect to be able to predict the exact judgment a subject will make on any given trial. And with a limited number of observations, we cannot expect that the actual proportion of Q responses will be precisely 0.6. Thus we can expect

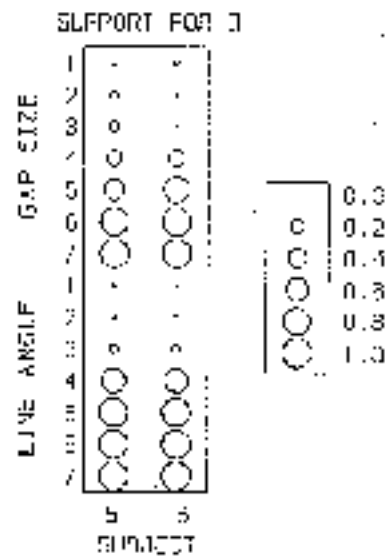

Figure 8.20
Plot of the FLMP parameter values indicating the degree of support for Q for the seven gap sizes and seven line angles for the two subjects shown in Figure 8.19. The scale on the right shows the relationship between area and parameter value.

some mismatch between the predicted and observed values, even if the model is correct. However, the observed variability should be equal to that expected from probability theory—in this case, binomial variability, which is the kind of variability observed in coin tossing. But it is possible to determine the expected binomial variability as a function of the observed response probabilities and the number of observations for each experimental condition. With this prediction of the expected variability we can ask if the fit of a model is poorer than what would be expected from random binomial variability.

As we noted earlier, the root mean square deviation (RMSD) can be thought of as a standard deviation. The standard deviation of a binomial distribution with two outcomes (e.g., heads or tails) is equal to the square root of its binomial variance, or

$$\sigma = \sqrt{\frac{pq}{N}} \tag{8.5}$$

where σ represents the standard deviation, and where p is the probability of one outcome (e.g., heads), q the probability of the other (where $q = 1 - p$), and N is the number of observations. Applying this equation to the Q-G task, p is equal to $P(Q)$, $q = P(G)$, and N is the number of times a

given experimental condition was tested for a given subject. For example, the prediction for a *Q* response might be 0.876 for a given condition. The prediction for a *G* response would be $1 - P(Q) = 0.124$. There were 12 observations for this condition, and the predicted standard deviation would be

$$\sigma = \sqrt{(.876)(.124)/12} = .095.$$

The benchmark RMSD is determined by computing the binomial variance (which is just the square of σ) for each of the 49 experimental conditions, averaging these 49 values, and taking the square root.

$$\mathrm{RMSD}(b) = \sqrt{\sum_{1}^{k}(pq/N)} \tag{8.9}$$

RMSD(b) is defined as the benchmark RMSD. In this equation, k is the number of experimental conditions (49 in the *Q-G* task), and $\sum_1^k$ means that we should sum the quantity in parentheses for each of the k pq/N values. These RMSD(b) values can be compared to the RMSD values from the fit of the FLMP to the observed results. The RMSD(b) values were 0.0687 and 0.0736 for the two subjects shown in figure 8.19. Although these benchmarks are somewhat bigger than the corresponding RMSDs, the differences were not significant when evaluated for all the subjects in the experiment. The benchmark comparison shows that the FLMP predicts the results about as well as any correct model can be expected to predict (see Wickens chap. 12, this volume, for a discussion of sampling variability.)

Up to this point, we have tested models against the distribution of ratings or the proportion of categorization judgments. Ideally, models should be capable of predicting several dependent measures. It turns out that the LNM and FLMP, with additional assumptions, can provide straightforward predictions of reaction time (RT) to make an identification judgment. I will not describe these extensions to the models here (the details can be found in Massaro and Hary 1986).

8.7 Context Effects in Pattern Recognition

The pattern recognition experiments discussed up to this point have manipulated one or two bottom-up sources of information, as in the *Q-G* task. Bottom-up sources refer to sources conveying information more or less directly from the stimulus, such as the gap size and line angle variables. Another source has been dubbed "top-down" to refer to other relevant information that a perceiver has available, including word knowledge and the information that emerges from the context accompanying the bottom-up sources. The context conveys information because a perceiver has prior knowledge about how the context constrains the bottom-up information.

I designed the following experiment because I was troubled by the large number of different theories of context effects in pattern recognition. It seemed to me that investigators too often devised studies to support only their own particular theories, rather than to distinguish among existing theories. My goal was to test among existing theories and to provide a challenging result for theories yet to come.

I adapted the logic of the factorial design of the *Q-G* task with two bottom-up sources to that of an experiment with a bottom-up and a top-down source. A bottom-up source was manipulated as in the earlier two experiment by varying the length of the horizontal line feature (see figure 8.6). To the extent the line is long, there is good bottom-up visual information for an *e* and poor visual information for a *c*. The goal was to independently manipulate a top-down source.

Consider one of the *ce* stimuli presented as the first letter in the context *-ofr* or in the context *-ell*. These different contexts provide top-down information. Only *e* follows English spelling patterns in the first context because the three consecutive vowels *eof* would violate English orthography. Only *e* is admissible in the second context because the initial cluster *cl* is an inadmissible English pattern at the beginning of a word. In this case, the context *-ofr* favors *c*, whereas the context *-ell* favors *e*. The context *-fss* and *-ssf* can be considered to favor neither *e* nor *c*. The first is an inadmissible context for *c* and *e*, and the second is admissible for both *c* and *e*.

My actual experiment combined 6 levels of line length with these 4 levels of context in a factorial design, giving a total of 24 experimental conditions (Massaro 1979). Each test letter was presented at each of the 4 letter positions in each of the 4 context levels. Figure 8.21 gives the 96 test items used in the experiment. A test item was presented for a short duration followed by a masking stimulus comprised of random letter fragments. Subjects were instructed to indicate whether they saw a *c* or an *e* in the test display.

The results of the experiment are shown in figure 8.22. Performance is necessarily worse than what we observed for single letters (shown in the right panel of figure 8.9) because in this context experiment, the test displays were presented for a much shorter duration and were also followed by a masking stimulus. As can be seen in figure 8.22, both the test letter and the context influenced performance in the expected direction. The proportion of *e* judgments increased with increases in the horizontal line length of the stimulus letter. The proportion of *e* judgments was largest

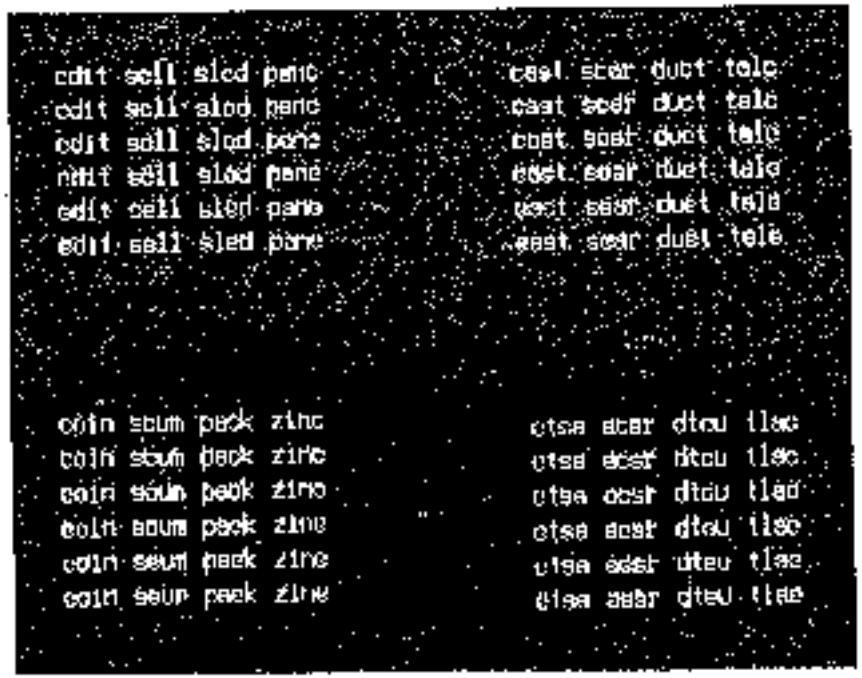

Figure 8.21
The test items used in the context experiment varying letter information and context (after Massaro 1979). The top left panel gives the test items for condition NA (neither admissible). The top right panel gives the test items for condition BA (both admissible). The left and right bottom panels give the test items for conditions CA (c admissible) and NA (neither admissible) respectively.

for the BA (both admissible) and smallest for the CA (c-admissible) context. Given the difficult display conditions, it is not surprising that there were also context effects for the unambiguous s and c letters. The NA (neither admissible) and BA (both admissible) conditions were intermediate and equal to one another. Further, the effect of context is larger for the more ambiguous test letters. (Although the curves might appear to be parallel, this is an illusion because the spread between the top and bottom curves is about twice as large in the middle of the curves than at the extremes.)

These results can be used to test our model contenders. Both the DFM and the FLMP can be extended by treating context as an additional feature or source of information. This top-down source would then be processed in the same manner that each model proposes for bottom-up sources. But recall that the DFM predicts additive effects of the two

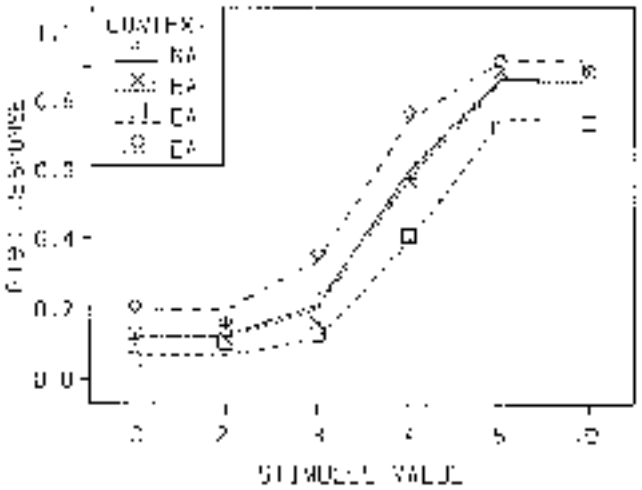

Figure 8.22
Observed (points) and predicted (lines) probability of a decision that an s is a function of the stimulus length of the test letter and the context (after Massaro 1979). Predictions of the FLMP.

information sources in parallel fractions for a plot of their effects, unlike the American football curve shape in figure 8.22. Given this fundamental mismatch between the observations and predictions, I will not subject the DFM to any further pain. The FLMP stands a much better chance of predicting the results, and I proceed with its quantitative test.

8.7.1 Test of the FLMP

Given that I assume that a top-down source functions in the same manner as a bottom-up source, the application of the FLMP to the results shown in figure 8.22 is exactly analogous to the Q-C experiment. Two independent sources of information are available at the evaluation process, and the subject evaluates the support for the alternatives c and s from each source. The value of the bottom-up visual information, denoted by s_i, is a number between 0 and 1 that indicates how much c is supported by the visual test letter. The value of the top-down information, denoted by c_i, indicates how much c is supported by the context. Given just two choice alternatives, it can be assumed that the support from a source for s is equal to one minus the amount of support for c. The integration operation multiplies the support from the two sources to give

$$M(c) = s_i c_i \tag{8.10}$$

$$M(s) = (1 - s_i)(1 - c_i). \tag{8.11}$$

At decision, the RGR operation gives the predicted probability of an e response, $P(e)$:

$$P(e) = \frac{a_i c_j}{a_i c_j + (1 - a_i)(1 - c_j)} \tag{8.12}$$

The lines in figure 8.22 represent the predictions of the FLMP. The FLMP was fit to the average results by estimating 6 values of a_i corresponding to the 6 levels of the horizontal line, and 4 values of c_j for the 4 context levels. The values of a_i, which increased systematically with increases in line length, were .12, .14, .22, .59, .86, and .86. The value of c_j was .66 for the e-admissible context and .32 for the a-admissible context. The two neutral contexts gave values of c_j close to .5 (.51 and .49 for e admissible both and e-admissible neither, respectively). The RMSD between the predicted and average observed results was only .024. The observations and predictions again have the signature of the FLMP in that the influence of one source of information is larger to the extent that the other source is ambiguous. This characteristic of the results was also observed in the Q-G task.

8.7.2 Sentence Context in Word Recognition

Many experiments have been carried out to demonstrate that sentence context influences word recognition. For example, in a classic study, Tulving, Mandler, and Baumal (1964) combined 8 exposure durations with 4 sentence context lengths in a word recognition task in which a briefly presented word followed the reading of a sentence context. One of the 18 sentence contexts was "Her closest relative was appointed as her legal ..." The test word, which you may have guessed, was "guardian." Subjects read either the last 0, 2, 4, or 8 words of the context part of the sentence, and the test word was then presented for either 0, 20, 40, 60, 80, 100, 120, or 140 milliseconds. Subjects were instructed to write down the test word and to guess if they were not sure of their answer. The left side of figure 8.23 presents the probability of a correct response as a function of the duration of the test word presented without context. As expected, accuracy increased with increases in word duration. Performance also improved with increases in the number of words of sentence context as shown in the middle of figure 8.23 for the 4- and 8-word context conditions. The rightmost part of figure 8.23 shows how well subjects could guess the test word using only the context information.

There have been few models of how such top-down context effects work, but analogous to the predicted context effects in the c-e experiment described above, the FLMP predicts that sentence context should improve word recognition. Consider the case of a 40 millisecond presentation and

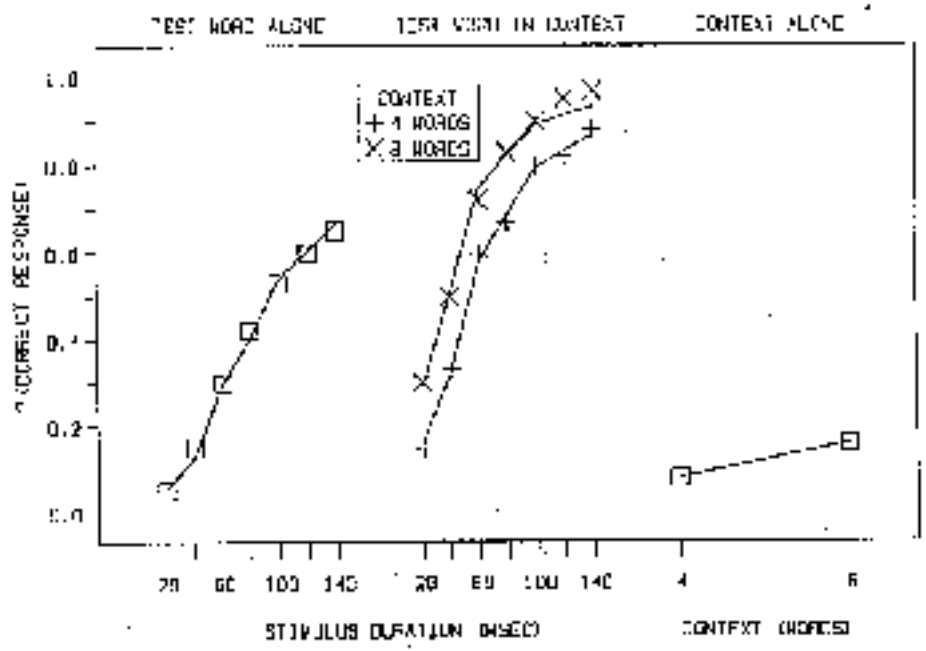

Figure 8.23
Observed (points) and FLMP (predictions lines) of the proportions of correct responses for the test word alone (left graph), test word in context (center graph), and context alone (right graph), as a function of stimulus duration and number of words in the context sentence (after Tulving, Mandler, and Baumal, 1964).

a 4-word context. When both of these sources are presented, figure 8.23 shows that accuracy was about 30 percent, larger than the 17 percent or the 10 percent accuracy given either the test word or the context alone. The lines in figure 8.23 show the predictions of the FLMP using the same 12 parameter values (11 levels of display duration, and 4 levels of context) for all conditions.

Given the success of the FLMP in predicting these and other results, it is challenging to find other models that can compete. To this end, I now develop and test several alternative models, based on the concept of neural networks.

8.8 Artificial Neural Network Models

The human brain contains billions of individual nerve cells called "neurons." Each neuron sends excitatory or inhibitory messages to other neurons, and the activity level of a neuron is determined by the approximately

1,000 excitatory and inhibitory signals it receives from other neurons. So far as we know, these neural messages are the basis of consciousness and thought (see Flanagan and Dryden, chap. 6, this volume). Recently, there has been a revived interest in constructing information processing models based on our understanding of the principles of neural mechanisms. Because they provide some novel alternative solutions and may also offer better descriptions of what the brain is actually doing than more traditional models, such models are important candidates for experimental tests.

In these models, scientists write computer programs to simulate the functioning of neurons and their interactions. Some computer programs try to describe the actions of only a few neurons, while others attempt to model the activity of thousands of neurons. But these programs provide only a general and abstract description of how real neurons might work, and at this stage of research, the models they generate are far removed from the complexity of the human brain. Nonetheless, scientists have learned a great deal from such explorations about how mental processes might occur.

In these simulation models, information is assumed to exist in the form of activations and inhibitions of neurolike units. All knowledge in the simulated neural system is contained in the connections among the units and the operations that map the input into the output. The units interact with one another via connections among the units, and therefore, such models are often referred to as "connectionist models." The connectivity is implemented by positive and negative weights among the units. As noted above, a weight is just a number that represents how the activity of one unit affects another unit. For example, a weight of 0 between two units, *a* and *b*, means that the two units are not connected and do not affect each other. On the other hand, a weight of +1 from *a* to *b* means that the output of *a* has a positive or excitatory influence on unit *b*. In a simulation program, the output of *a* is multiplied by +1, and this value becomes an input to *b*. Finally, a negative weight, such as −1, between *a* and *b* represents an inhibitory connection. The output of unit *a*, multiplied by the negative weight, is subtracted from the sum of all the other inputs to unit *b*. The units in connectionist models are generally organized into several lists or "layers," such as an input and an output layer. This connectionist framework has been used to model a broad variety of psychological phenomena.

9.8.1 Connectionist Model of Perception

I begin with a connectionist model of perception (CMP) for the Q-G discrimination task I described in section 9.6.2. I will assume that for this task, there are 14 input units: 7 line angle and 7 gap size feature detector units,

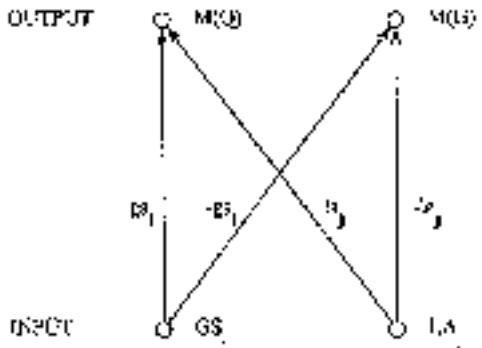

Figure 8.24
Illustration of the connectionist model with two of the four input units, GS_i and LA_j, and two output units, Q and G. The input units shown correspond to a gap size and a line angle feature detector, and the output units correspond to the two list alternatives, gs_i and la_j represent the weighted outputs of the gap size feature detector, and la_i and la_j are the outputs of the line angle feature detector. $M(Q)$ and $M(G)$ are the outputs of the logistic transformation of the summed inputs to Q and G, respectively.

corresponding to the 7 levels of the line angle and gap size features in figure 8.14, respectively. I will also assume that each of the 7 levels of line angle in the stimulus activates only one of the 7 line angle units. Or, to put it another way, I am assuming that there are 7 line angle feature detector units, each of which responds to only 1 of the 7 levels of line angle in the stimulus set. Similarly, I assume that there are 7 gap size detector units. In addition to these 14 input units, this CMP has 2 output units, whose outputs, $M(Q)$ and $M(G)$, represent the degree of support for Q and G, respectively. Each of the 14 input units is connected to each of the 2 output units. Figure 8.24 illustrates the CMP for 2 of the 14 input units and both output units.

The activation level of an input feature detector unit is zero unless it is activated by a test stimulus feature, when its value becomes one. The activation of an output unit by an input unit is given by multiplying the input unit's activation and a weight w. With a particular test stimulus, there will be two active input units, GS_i and LA_j, corresponding to the active gap size and line angle feature detectors, and the activation entering output unit Q is $gs_i + la_j$, where $gs_i = w_iGS_i$ and $la_j = w_jLA_j$. Similar to the Pandemonium model described earlier, the weight that connects an input feature detector to an output unit represents the support that the feature detector provides for that output when the feature detector is active. From this, it can be assumed that activation entering output unit G from each input unit is simply the same as the value entering Q but with opposite sign. Thus the activation entering output unit G would be $-gs_i - -la_j$, with $gs_i = -w_iGS_i$ and $la_j = -w_jLA_j$. Given that GS_i and LA_j are

either 0 or 1, $gs_i = s_i$ and $la_j = a_j$. The output of an output unit is given by the sum of the two input activations, passed through a logistic function. Because real neurons can only respond up to some maximum output, a logistic function serves to map a large range of input activation values into output values from near 0 (for large negative activation values) to near 1 (for large positive activation values).

The resulting output of the Q output unit after applying the logistic function to the unit's activation value gives the support for alternative Q as shown in equation 8.13. In this equation, $(gs_i + la_j)$ represents the output from the two active feature detectors to the Q output unit.

$$M(Q) = \frac{1}{1 + e^{-(gs_i + la_j)}} = \frac{1}{1 + e^{-(s_i + a_j)}} \tag{8.13}$$

A similar equation (equation 8.14) gives the support for G.

$$M(G) = \frac{1}{1 + e^{-(gs_i + la_j)}} = \frac{1}{1 + e^{-(s_i + a_j)}} \tag{8.14}$$

Once the support for Q and G have been computed, the response can be selected using the same RGR rule that I used before, as shown in equation 8.15.

$$P(Q|GS_i, LA_j) = \frac{M(Q)}{M(Q) + M(G)} \tag{8.15}$$

This connectionist model can be tested against the results in the same manner as the FLMP. The 14 weights are now the free parameters that are adjusted to maximize the fit of this CMP to the Q-G results shown in figure 8.19.

You might be surprised to learn that this specific CMP is mathematically equivalent to the FLMP formalized in the previous section. I was certainly amazed when I first discovered and proved this equivalence, and it taught me that two very differently formulated models might sometimes make equivalent predictions. Given this equivalence, the FLMP predictions in figure 8.19 are also predictions of the CMP (Massaro and Friedman 1990). After this experience, I am more open to looking for similarities among different models, as well as differences. Although the CMP and FLMP are not identifiably different for this Q-G task, they do make different predictions in tasks with three or more response alternatives, and the FLMP has been shown to provide a more accurate description of pattern recognition tasks with four alternatives (Massaro and Friedman 1990).

8.8.2 Interactive Activation Model

To learn more about artificial neural networks, I will now consider a well-specified and influential neural network model of word recognition. This

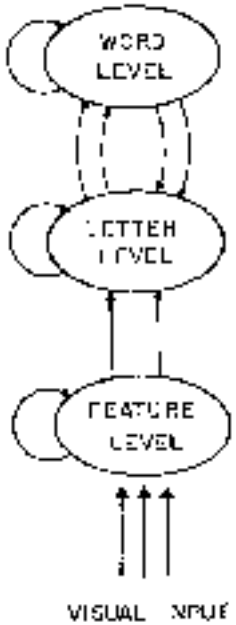

Figure 8.25
Three levels of units in the interactive activation model (IAM) from McClelland and Rumelhart 1981. Each level represents a set of units of the specified type. Activation connections are signified by lines with an arrow, whereas inhibitory connections are signified by lines with a dot. Visual input activates units at the feature level, which inhibit each other and activate the units of letters that are consistent with these features and inhibit the units of letters that are inconsistent: these features. Similarly, the activation of letter units activates the units of consistent words and inhibits the units of inconsistent words. The interactive activation aspect of the IAM involves activation from the word level to the letter level. The activation of a word unit is assumed to activate the units of its component letters. In addition, activated word units inhibit other word units.

particular model was designed to account for context effects in letter and word perception (McClelland and Rumelhart 1981). The interactive activation model (IAM) makes an important assumption that distinguishes it from models like the FLMP and the CMP. The IAM assumes two-way connections among some of the units, as opposed to the one-way flow of information from input to output assumed in the FLMP and the CMP. In the IAM there are three layers of units: features, letters, and words; the connections among the units across the different layers are shown in figure 8.25.

Presentation of a stimulus word to the network activates features corresponding to stimulus input. These active feature units then activate the units of letters that are consistent with these input features and inhibit the

units of letters that are inconsistent. Similarly, the active letter units then activate the units of consistent words and inhibit the units of inconsistent words. The interactive activation aspect of the IAM involves activation from the word layer to the letter layer. An active word unit activates the units of its component letters. In addition, an activated word unit inhibits other word units. Thus the feature detection units are first activated by the stimulus, then the letter units receive activation from feature units and, after the letter units activate some word units, also obtain activation from the activated word units. An important characteristic of the IAM is that all words that share letters with the presented word will be activated to some degree, and will then activate their component letters.

There are several major differences between the IAM and CMP that require explanation. The CMP generates its output with just one pass of information from input to output. We call this a "processing cycle." That is, a stimulus activates the input units, which in turn activate the output units, which make information available to decision. In the IAM, on the other hand, there are several processing cycles to permit the units to interact before a response is chosen. In addition, there is no sharp distinction between input and output units in the IAM. Depending on the task (e.g., letter or word identification), the response can be based on the active units in any layer. McClelland and Rumelhart (1981) assumed that if subjects were asked to report what letters were present, the response was determined by activation at the letter level.

The IAM was designed to account for context effects in word recognition, such as the finding that context letters influenced letter recognition in the c-e study. The IAM account of context effects explains the benefit of the context solely in terms of the contribution of the knowledge of specific words in the subject reader's lexicon, rather than from general knowledge of English spelling rules. Activated word units send activation to their component letter units and inhibition to other letter units.

Let us consider an ambiguous letter presented in the context *-dit*. Activation of features and letters by the context will activate the word unit *edit*. After several processing cycles, this word unit will have activated the letter unit corresponding to the letter *e* in the initial position. Thus, the subject will be more likely to report *e* than *c*, given this context. Interactive activation explains this context effect in terms of activation from the word layer to the letter layer. Like the other models we have considered, it is the relative amount of activation at the letter level that is critical for performance in this task (i.e., the RGR is used in the decision operation).

Figure 8.26 illustrates the time course of the letter and word units that are activated when the word *work* is presented. In this example, it is assumed that all of the features of the first three letters are successfully

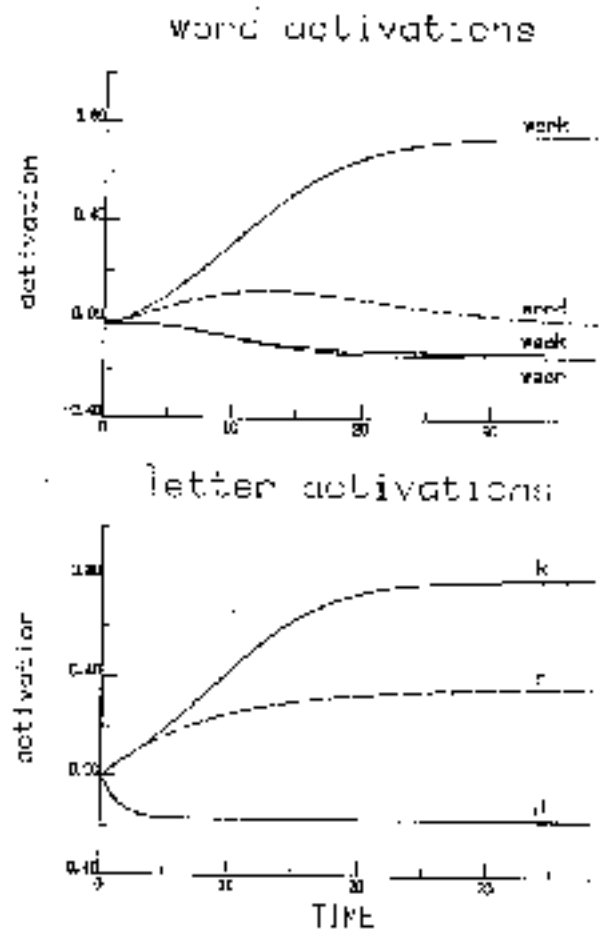

Figure 8.26
The time course of activations of selected units at the word and letter levels (from McClelland and Rumelhart 1981)

detected along with features consistent with the letters *k* and *r* in final position. The bottom graph shows how the *k* and *r* letter units initially show the same increase in activation. Because activation first goes from feature to letter to word layers, the influence of activation at the word layer (top graph) on the letter layer is necessarily delayed relative to activation from the feature layer to the letter layer. Given that the feature information is equally consistent with both *r* and *k*, it is the activation from the unit *work* in the word layer that is responsible for the appearance of increased activation of the *k* unit relative to the *r* unit in the bottom graph.

One difficulty in comparing the IAM and FLMP is the somewhat different procedures that are used in testing mathematical and more complex simulation models. Mathematical models, such as the FLMP, are formulated as equations that allow us to estimate parameters to obtain optimal prediction of a set of observations. A measure of goodness of fit, such as RMSD, can then be used to evaluate the model. Simulation models, on the other hand, do not have a set of equations that predict observations directly. Instead, tests of simulation models, such as the IAM, have usually been carried out using computer simulation. Computer simulation is also an excellent method of testing theories because, like the simpler computational models we have considered, they must be precisely specified. The assumptions of the model are written in the form of a computer program that mimics the hypothetical behavior of the subject. The program is then given a set of test trials, and the output of the model is compared to a subject's performance.

Implementing the IAM simulation requires specification of the features, letters, and words in the three levels of units. In the original simulations, the model assumed a particular set of feature units, 26 letter units, one for each letter of the alphabet, and a set of 1,179 four letter word units. Each of the 26 letter units has an excitatory or inhibitory connection to each of 1,179 four-letter word units, and each of these 26 × 1,179 connections, as well as connections from the feature layer and the the other connections that are required by the model must be specified by weight parameters. Even if simplifying assumptions are made (e.g., all the excitatory connections from letter units to the relevant word units have the same connection weight), there are many parameters to be estimated. Rather than estimating the parameters, a particular set of parameter values is chosen, and the model is run through a number of processing cycles in order to determine the predicted results. This aspect of a simulation model takes substantial computer time, and each new set of parameters requires another simulation run. Thus many simulation runs are generally required, and finding a good set of parameters is very time-consuming. For this reason, simulation models rarely specify their parameters in advance, and simulated performance cannot be expected to match actual performance as closely as the fit of a mathematical model. The goal is to see if the simulated results have the same general form as the observed results, rather than matching the results exactly.

In the *c–e* study the IAM correctly predicts that subjects will be more likely to report *c* in the context *-ell* than in the context *-oin*. The model also predicts intermediate performance for the *-ion* and *-ell* contexts. Thus both the FLMP and IAM appear to be successful in describing the joint influence of bottom-up and top-down sources of information. More recently, several investigators (McClelland, 1991; Massaro and Cohen, 1991) have attempted to place the FLMP and IAM on more equal footing to allow more direct comparisons. One method that has been used is to reduce the complexity of the IAM by using miniature artificial neural networks so that optimal parameter values can be estimated. One network designed to predict context effects is shown in figure 8.27 (McClelland, 1991; Massaro and Cohen, 1991). Although simpler than the original IAM, the network is designed to explain how context could influence the recognition of *c* versus *e*.

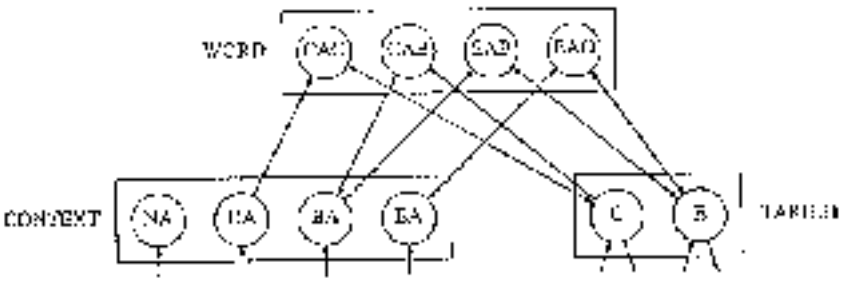

Figure 8.27
Network used in the simulation of the IAM model applied to the experiment of Massaro (1976). The inhibitory connections between units within the word, context, and target levels are not shown in the network. The target units are C(E) and E(C). The context units are NA (neither admissible), CA (*c*-admissible), EA (both admissible), and EA (*e*-admissible). The word units are CAO (*c*-admissible only), CAE (*c* admissible finish), EAO (*e* admissible finish), and EAO (*e*-admissible only).

Three sets of units are assumed: "target," "context," and "word." Target and context units are activated by the target (i.e., test) and context letters, respectively, analogous to the bottom-up and top-down sources of information. Activation of context units and target units activates word units. Consistent with the IAM assumption of interactive activation, all units within the context and target levels are bidirectionally connected to all word units. Only the excitatory connections are shown in the figure. Also, within each of the three sets of units, each unit also has inhibitory connections to all other units within that set.

The target units C and E are activated to varying degrees by a test letter. As in the FLMP, I will assume that increases in the length of the horizontal line will increase the activation of the E unit relative to the C unit. The context units are activated by the different contexts. The NA (neither admissible) unit would be activated by all contexts that support neither *c* nor *e*, such as the context *-tsa*. The CA (*c*-admissible) unit would be activated when the context supports only *c* and the EA (*e*-admissible) unit would be activated when the context supports only *e*. The BA (both admissible) unit would be activated when the context supports both *c* and

c such as the context *rest*. The word unit labeled "EAB" (e admissible both) corresponds to all words that have e in the critical position but also have word "neighbors" that are identical except for the letter e in the critical position. The word *east* would be an example, with the word *rest* as a word neighbor. Thus the context item *-est* would activate the EAB unit, and this activation would feed back and activate the target unit E. This same context would activate the CAB (c-admissible both) unit, and this activation would feed back and activate the target unit C. The word unit labeled "EAO" (e-admissible only) represents all words like *edit* and *debt* in which only e is admissible in the context. The word unit labeled "CAO" (c-admissible only) represents words in which only c is admissible, such as *coin* and *nice*. As can be seen in figure 8.27, the letter context activates the context units, the test letter activates the target units, and these units send their activation to the word units. Because of the interactive activation assumption, the word units in turn send their activation downward to the context and target units.

Although I now present the mathematical form of the IAM (McClelland and Rumelhart 1988), the reader can skip forward to the paragraph after equation 8.25 without sacrificing an understanding of the gist of the model. Basically, the equations formalize the following five operations for each unit: (1) each unit is set to some initial level of activation; (2) excitatory and inhibitory inputs from the target and context as well as from other units are computed; (3) the inputs are weighted and summed; (4) the change in activation is determined; and (5) the support for each response alternative is computed.

Initially, for each unit, i, its activation, act_i, is set to the "resting" level, $rest_i$, which represents the unit's activation when it has no input. Then, on each processing cycle of the model, for each unit, i, the excitatory input, ext_i, and inhibitory input, inh_i, are computed from the product of each sending unit, indexed with j, with its connection weight w_{ij} as follows:

$$ext_i = \sum_j \max(0, w_{ij}) \times (0, act_j); \tag{8.16}$$

$$inh_i = \sum_j \min(0, w_{ij}) \times (0, act_j), \tag{8.17}$$

where w_{ij} is the weight connecting unit j to unit i, $\max(x, y)$ represents the maximum of the two values, x and y, and $\min(x, y)$ represents the minimum of the two values. In this particular model, all weights w_{ij} are either +1 or −1, so that equation 8.16 adds up all the inputs with positive connections and equation 8.17 adds up all the inputs with negative connections. Activations less than 0 are ignored in these summations. Next, for each unit, i, the summed net input, net_i, is computed from the weighted

sum of ext_i, inh_i, and $extr_i$, where $extr_i$ represents input to unit i from the stimulus.

$$net_i = \alpha \times ext_i + \gamma \times inh_i + \mathit{rstr} \times extr_i, \tag{8.18}$$

where α is an overall weight on excitatory connections, γ is an overall weight on inhibitory connections, and rstr is the weight on external inputs.

Next, the change of activation for each unit for the upcoming cycle, Δact_i, is computed as follows:

$$\text{if } net_i > 0,\ \Delta act_i = net_i \times (M - act_i) - \mathit{decay} \times (act_i - \mathit{rest}_i); \tag{8.19}$$

$$\text{if } net_i < 0,\ \Delta act_i = net_i \times (act_i - m) - \mathit{decay} \times (act_i - \mathit{rest}_i), \tag{8.20}$$

where M is the maximum allowed activation, m is the minimum allowed activation, and decay is the rate at which each unit tends to return to the resting level. Then each act_i is adjusted by adding Δact_i:

$$act_i = act_i + \Delta act_i \tag{8.21}$$

Finally, each act_i is adjusted, if necessary, to remain in the interval m to M. The logic of this adjustment is analogous to that underlying the use of the logistic function in the CMP.

$$\text{If } act_i > M,\ act_i = M; \tag{8.22}$$

$$\text{If } act_i < m,\ act_i = m \tag{8.23}$$

McClelland (1993) uses the following set of parameters: $\alpha = .1$, $\gamma = .1$, $\gamma = .1$, $\mathit{decay} = .1$, $M = .1$, $m = -.2$, and $\mathit{rest} = -.1$. In the network shown in figure 8.27, it can be seen that the effects of the large letter and context are combined via the units in the word layer. The activations of word units are fed back to the target and context units, changing their activations in a manner that reflects the activations of both target and context units. In this manner, the joint effect of target and context are represented in the activations of units in both the target and context layers. This passing of activation from unit to unit occurs for a number of processing cycles. Then the activation, act_i, of a target unit is transformed by an exponential function into a strength value S_i, as shown in equation 8.24.

$$S_i = e^{kact_i} \tag{8.24}$$

The strength value S_i represents the strength of alternative i. The probability of choosing a particular alternative, PR_i, is based on the activations of all relevant alternatives, as described by the RGR,

$$P(R_i) = \frac{S_i}{\sum}, \tag{8.25}$$

where $\sum$ is equal to the sum of the strengths of all relevant alternatives. Massaro and Cohen (1991) set the constant k to 5.

To fit the simplest IAM to the c-e study in figure 8.21 requires 11 free parameters: 6 a target value inputs corresponding to the 6 levels along the c-e continuum (with the e target input equal to one minus the input to the c target unit), 4 context value inputs, and a parameter θ that gives the number of processing cycles at which a decision is made.

This IAM–RGR was fit to the observed data shown in figure 8.22 by minimizing the differences between the predicted and observed values. For each of the 24 experimental conditions (6 levels of the c-e continuum and 4 values of the context), a simulation trial was run with a set of parameter values, and a goodness of fit was computed. As in the fit of the FLMP, the parameter values were changed systematically to maximize the goodness of fit of the IAM.

Figure 8.28 gives the predictions of the IAM for the results of the c-e study. As can be seen in the figure, the IAM–RGR is not capable of describing the influence of the different types of context, and does a poor job describing the results, with a RMSD value of .067. The IAM is not

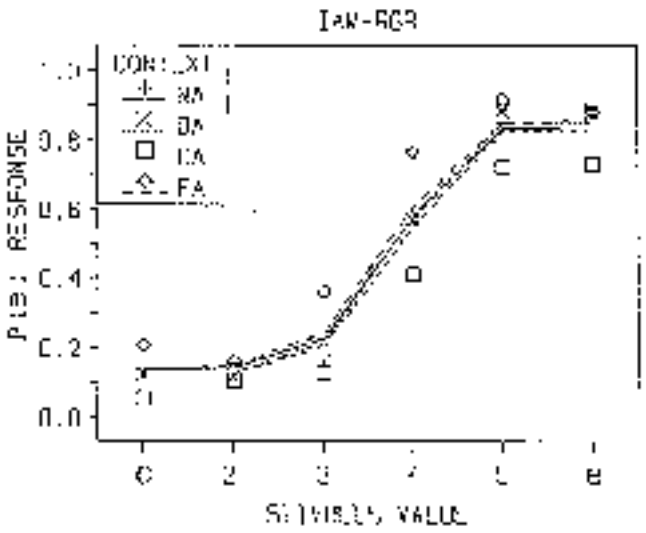

Figure 8.28
Observed (points) and predicted (lines) probability of an e response as a function of stimulus value and context. Predictions of the IAM–RGR model.

able to predict the type of interaction observed between stimulus information and context, and fines its best fit by predicting very little effect of context. Given its failure, it is only reasonable to revise the model to attempt to bring it into line with the results.

8.8.5 IAM with Input Noise and Best-One-Wins Decision Rule

McClelland (1991) placed the blame for the IAM's failure to predict these results on the decision stage of the model rather than on some other process, such as interactive activation. He modified the IAM by adding noise (random variation) to the input and by changing the RGR decision rule to a best-one-wins (BOW) rule in which the response alternative corresponding to the most active unit would always be chosen (McClelland 1991). With these two modifications, the predictions of IAM appeared to be consistent with the empirical observations (and the predictions of the FLMP).

To provide an empirical test of the new IAM, simulation trials were run by adding noise to both the target and context input values. For each of the simulated trials, a BOW decision was made on the final target activations. The simulated trials give the predicted proportion of e responses for each of the experimental conditions. Figure 8.29 shows the fit of this 11-parameter IAM. The fit of this model did not improve on the fit of the

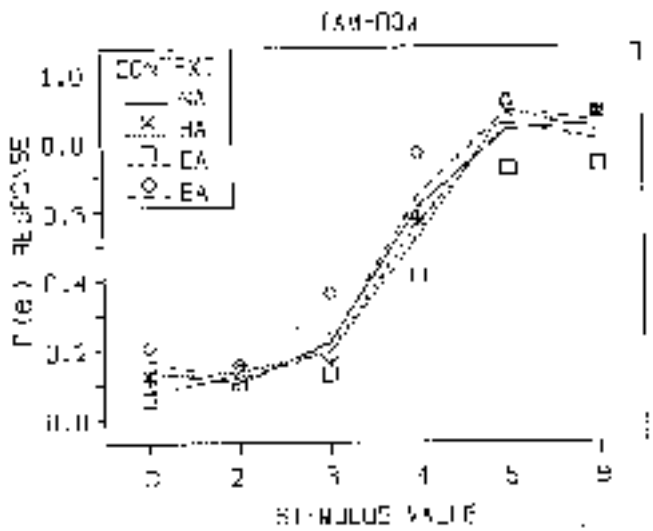

Figure 8.29
Observed (points) and predicted (lines) probability of an e response as a function of stimulus value and context. Predictions of the IAM–BOW model.

deterministic IAM with the RCR decision rule. The RMSD obtained for this model was .056, over twice that found for the FLMP.

A second, 12-parameter IAM was run, starting with the parameters of the 11-parameter model, but with the amount of noise added to the input as a free parameter. No improvement was seen in the RMSD. Thus we can conclude that these results support the FLMP over the IAM. Other results agree with this conclusion by demonstrating that interactive activation appears to be an inappropriate process to account for the influence of context on perception (Massaro and Cohen 1991). One caveat, however, is that the experiments have falsified only specific implementations of a neural network theory of interactive activation. It is always possible that some other implementation will be able to account for the results. (See Anderson, chap. 7, this volume, for another perspective on connectionist models.)

The reader who has persisted to this point has received at least a taste of firsthand experience in the endeavor of computational modeling. The most exciting aspect of this enterprise is the challenge to formulate and test your ideas about mind and behavior. There is no shortage of good ideas for models and experiments. The participant in computational modeling can expect to be constantly engaged by it.

8.9 Justification of Computational Modeling

Computational modeling is the formal, quantitative description of behavior by the interaction of a set of simpler component processes. We attempt to understand the complexity of everyday behavior in terms of simpler mechanisms working in combination. Computational modeling necessarily describes the processes causing behavior and not just the behavior itself. Computational modeling is a valuable approach because it helps to overcome difficult problems in psychological inquiry. I briefly describe these problems and then how they are made easier to solve by computational modeling.

8.9.1 Difficulties in Psychological Inquiry

The first difficulty is that behavior is complex. The fact is that complexity is endemic in nature generally. Why can we walk on the moon and put a camcorder in the helmet of a football player, but not predict the weather? One reason is simply the many and varied influences on the weather (temperature patterns, solar radiation cycles, moisture, ocean currents, seismic events, and so on). A related reason is the butterfly effect: small differences at one point in time may have dramatic consequences at some later time. Because the weather is critically sensitive to slight changes in

initial conditions, a butterfly over the North Pacific, in at least a 747, might influence the weather in Santa Cruz a few days later. This butterfly effect is a surprising discovery in the new field of chaos theory (Gleick 1987) and is captured in the folklore example:

> For want of a nail, the shoe was lost;
> For want of a shoe, the horse was lost;
> For want of a horse, the rider was lost;
> For want of a rider, the battle was lost;
> For want of a battle, the kingdom was lost.

Of course, complexity is the hallmark of causes of behavior as well. Each of us has had uniquely different histories and, more often than not, would behave differently in the same circumstances. Freud probably overemphasized the importance of our early childhood, but he might have been right about how some early experiences could have dramatic consequences later in life. If behavior is dependent on many influences and on our distant pasts, it will be difficult to uncover simple behavioral laws.

A second difficulty in psychological research is that we tend to interpret the world as more orderly than it actually is. Suppose I tell you that I have in mind a rule for producing a sequence of numbers, and I ask you to try to guess the rule. If you are told that the sequence 2, 4, 6 was generated by this rule, you will probably guess that the rule is "The numbers increase by two." In fact, this guess would be much more specific than the rule I have in mind. We naturally tend to see the rule as more specialized than necessary. As scientists, this bias to interpret situations too simplistically can impede understanding and accurate prediction.

A third difficulty is that people have a bias to validate their beliefs, a confirmation bias. People will tend to test the hypothesis "The numbers increase by two" by generating sequences like 16, 18, 20, or 1, 3, 5, and so on—failing to see that the positive outcomes are not only consistent with this rule but also with a simpler rule, "The numbers increase." This type of bias also occurs in scientific inquiry. Given some observations and an explanation of them, scientists also display a confirmation bias as they engage in hypothesis testing. A scientist might develop the hypothesis that such and such should occur given such and such situation. It is only natural that the same researcher will be motivated to test the accuracy of this hypothesis, and it is well documented that scientists, like all humans, actively search for evidence to support their beliefs, often ignoring alternative hypotheses.

Leo Tolstoy rationalizes the human nature of confirmation bias as follows:

> I know that most men, including those at ease with problems of the greatest complexity, can seldom accept even the simplest and most

> obvious truth if it be such to oblige them to admit the falsity of conclusions which they have delighted in explaining to colleagues, which they have proudly taught to others, and which they have woven, thread by thread, into the fabric of their lives. (as quoted in Ford 1986, 7).

Confirmation bias even occurs in the most seasoned scientists. Although less common, there is even evidence of downright cheating in scientific inquiry, in addition to the more acceptable everyday practices that sometimes maintain our allegiance to shaky theoretical perspectives.

A fourth difficulty in psychological inquiry, which may be understandable, given the first three, is that there are several alternative theories to describe most phenomena. For example, there is currently a great deal of controversy concerning language acquisition and use. At issue is whether a child's productive ability requires an internalization of rules of grammar, or whether simpler learning mechanisms are sufficient. In a classic experiment, young children were able to generate plurals of pseudowords they had never heard before. The obvious interpretation at that time was that the children used a rule to achieve the "correct" outcome. There is a rule for forming plurals and the children may use this rule to guide their behavior. On the other hand, the children might have performed correctly simply by using analogies to specific words that they already knew. In this case, a child who knows the plurals of rug, bug, and mug could induce from these instances that a good guess for wug might be wugs. Although generating wugs based on analogies might be regarded as a good example of inductive generalization, it would not be a grammatical rule.

In summary, we face the complexity of behavior, we tend to impose more order than may actually exist, we display a confirmation bias in testing hypotheses, and we are overwhelmed with several conflicting theoretical explanations. These difficulties can be used to argue for a particular form of scientific practice.

8.0.2 Implications for Psychological Inquiry

Four features of the computational approach can help to address these four problems. First, the computational approach enables the investigator to develop highly specific, precise, and simplified experimental studies in order to reveal fundamental regularities or laws related to the phenomena of interest. The idea is that these regularities or laws cannot be revealed or tested in highly complex situations because of the many influences at work that will not be easily disentangled. A valuable theory might be capable of predicting behavior in the laboratory but not in the naturally varying environment, where the complexity of the natural setting might far exceed what any scientific theory could hope to predict. Theories will have predictive power only in the laboratory in which the complexity of everyday life can be simplified, measured, and controlled. One should not become disheartened, however, because these same theories help us understand and explain complex situations. But complex situations do not usually permit definitive tests of competing theories.

Second, the computational approach enables the investigator to perform a fine-grained analysis of the results to distinguish among competing models. Only by thorough, systematic, and even paranoid (there is always another possible explanation lurking in the wings) analyses can an investigator eliminate alternative models. And only this elimination will reduce the set of models consistent with the observations of human performance. Small deviations between predictions and observations must be taken seriously. The German astronomer Johannes Kepler was convinced that the tiny, but systematic, eight minutes of arc deviation between the actual orbit of Mars and that predicted by Ptolemaic theory was meaningful. Taking this result seriously perhaps contributed to Kepler's patience during the thirty years of inquiry it took to arrive at his three laws of planetary motion.

Third, the computational approach enables the investigator to test between alternative models of performance using the research strategy of falsification and strong inference. Falsification has the goal of actually falsifying a model, and strong inference seeks to distinguish between two or more different models. Within the computational framework, the investigator should develop opposing models and devise experimental situations that allow tests between the predictions of these different models. Simply accumulating evidence that is consistent with a given model is not an ideal strategy because the data might be equally consistent with a diametrically opposed model. That is, most results are indiscriminately confirmatory: they support a variety of hypotheses. For example, the positive effect of context found in the [illegible] study (figure 8.2.2) is consistent with all theories that predict a positive context effect. Progress is made only when alternatives are eliminated, as in the case of falsifying the IAM account of context effects. Furthermore, it must be remembered that elimination disqualifies a theory—even if it is consistent with many other findings. The idea of whole-word shape being functional in reading is a good example. Although this old idea was consistent with an abundance of research, it was disqualified only when experiments were motivated by falsification and strong inference. For example, the advantage of recognizing letters in words also occurs when the words are written in mixed upper- and lowercase, which should eliminate any benefit from word shape.

And fourth, the computational approach enables the investigator, ever on guard against unfalsifiable theories, to permit only models with "discriminating taste" to survive. Discriminating taste means that a model

must be specific enough to predict actual results, not the universe of possible results. The investigator therefore tests the theory not only against observed results but also against a range of hypothetical results that are known not to occur.

Good scientific inquiry dictates that we actively attempt to eliminate alternative models. However, that is difficult to do with theoretical frameworks whose flexibility or "superpower" enables them to predict a wide range of alternative results; this appears to be the case for physical symbol systems and connectionism. In both of these approaches, the time course of the component processes of a model does not usually match the processes that occur in human beings. Therefore, the observed data often will not be sufficient to decide among alternative models. Furthermore, although certain models within both of these theoretical approaches can be falsified, other models can be devised to reproduce any possible result. This makes these theoretical frameworks unfalsifiable at a general level.

8.10 Metatheoretical Issues and the Computational Approach

There are also two metatheoretical issues that must be addressed to help situate the computational approach. The first issue, identifiability, has to do with the feasibility of discriminating among alternative possible models or explanations of a particular set of behaviors. The second concerns the question of whether human pattern recognition is optimal.

8.10.1 Identifiability Issue

Identifiability concerns whether or not experimental results can decide among different models. A lack of identifiability means that the predictions of one model can always be mimicked by another model—precluding any definitive test between the two models. As an example, there are two well-known methods of addition: successive addition and a lookup table. With successive addition, 5 + 3 is computed by adding 1 to 5 three times. A lookup table would simply have the entry "8" stored for this problem. These two methods can be viewed as two different psychological processes. The input-output functions do not allow us to distinguish between these two models of addition: the same input-output functions are observed for both. The general point is that the predictions by one model of an experimental result are not necessarily unique. Some other model can be manufactured to give exactly the same predictions.

Scientific inquiry can potentially choose among equally accurate models by extending the empirical database, by evaluating the models on the basis of parsimony, and by testing among models using the principles of falsification and strong inference described earlier. Extending the database is a valuable strategy for distinguishing models that make identical input-output predictions. To return to the addition example, reaction time (RT) is a valuable dependent variable. In an illustrative series of experiments on how children add two numbers, it has been possible to distinguish between two viable models of addition by measuring RTs to different problems. The results indicate that 6 + 5 takes about the same amount of time as the problem 4 + 3. However, both these problems take longer than 7 + 1. In total, the results indicate that, at one stage of development, the child recognizes the numbers, chooses the larger one, and then adds the smaller number by counting from the larger to the smaller in steps of one (Groen and Parkman, 1972; see also Anderson's discussion of arithmetic learning, chap. 7, this volume).

Parsimony, or the simplicity with which a model is formulated, can also be used to choose among models. In general, if we have two models that make the same predictions, we should prefer the simpler model. We saw that a template-matching model was not parsimonious because we were forced to assume that there could be a different template for every input; as developed, this was not a viable model of letter recognition. Finally, falsification and strong inference guide the experimenter to new situations that allow tests among previously nonidentifiable models. In summary, we have seen that identifiability is not an insurmountable problem.

8.10.2 Optimality of Pattern Recognition

A perennial issue in psychology is whether human decision making and choice are optimal. Research during the last four decades has consistently found that human judgment does not appear to be perfectly rational or optimal. These findings and others widened the gulf between experimental psychology and other disciplines such as classical economics which take optimality and rationality as first principles of human behavior.

We have seen that the FLMP gives a good description of pattern recognition. If indeed pattern recognition is central to much of our perceptual and cognitive behavior, we might speculate that we have evolved to recognize patterns in an optimal manner. Thus I was pleasantly surprised when I discovered that two aspects of the FLMP, having continuous information at evaluation and following a multiplicative integration rule, can be shown to be optimal. Discrete information or an additive type of integration is necessarily nonoptimal.

The relative goodness rule (RGR) assumed by the decision stage might be considered to be nonoptimal, however. According to this rule, subjects do not always choose the strongest candidate, in essence, they probability match the value given by the RGR. This decision rule might be considered nonoptimal because it does not maximize the accuracy of

performance. For example, if the RGR gives 0.8 for an alternative, that alternative is chosen 0.8 of the time in probability matching.

Another possible rule, called "maximizing," would be to always choose the alternative with the highest relative goodness because this alternative is necessarily the most likely one to be correct. Suppose that on a particular trial, the RGR gives 0.8 for a particular alternative, and the probability that this alternative was actually presented is also 0.8. Then, with the RGR, the probability of being correct on these trials would be 0.8 × 0.8 = .64. But with the maximizing rule, the probability of being correct would be 0.8 × 1 = 0.8. As pointed out by Massaro and Friedman (1990), however, a subject's goal might be to communicate the relative goodness of match rather than to maximize accuracy. Because a subject is required to respond with a discrete alternative, the only way this can be done is by probability matching the output of the RGR.

Although performance is well described by the FLMP, which assumes (by the RGR) that subjects probability match, it turns out that a somewhat different model, which assumes subjects always maximize (choose the best matching alternative), can also describe the data. Here we have an example of an identifiability problem. This alternative model arises if there is random variability (noise) in the output of the integration process. In this case, the subject always chooses the best-matching alternative at decision but, because of this noise, the alternative that always gives .9 in the absence of noise sometimes gives a lower value than another alternative. As a result, subjects sometimes respond with the other alternative. Thus subjects are behaving optimally by maximizing, but final performance is nonoptimal because of noise. It is interesting to note that this is precisely the assumption that has been so successful in an approach to human decision making called "signal detection theory" (see Swets, chap. 13, this volume, for an account of this theory.) Thus the maximizing rule with noise and the RGR make very similar predictions. Our current results cannot tell which of these two decision rules is used.

Note

The research reported in this chapter and the writing of the chapter were supported, in part, by the Public Health Service grant PHS 301 N2 20814 and National Science Foundation grant BNS 88-12728. The author thanks Bill Ferrer, Joel Norman, and Stephen Pitkin for help in commenting on an earlier version of the chapter.

Suggestions for Further Reading

Suggested readings on the topic of reading are an edited volume by Willows, Kruk, and Corcos (1993) [illegible] of the journal *Experimental Psychology: Human Perception and Performance* [illegible] "A variety of information-processing [illegible]" by Massaro and [illegible] [illegible] (1992) [illegible] *Cognitive Psychology* [illegible] Gibson (1969).

Problems and Questions for Further Thought

8.1 Assume a two factor factorial design with three levels of both variables. Given parameter values of .1, .5, and .9 for one factor and .7, .5, and .9 for the other factor, work out the predictions for both the DPM and the FLMP.

8.2 Show how it is essential to assume continuous information in the FLMP by deriving predictions for the design in problem 8.1 when the parameter values are 0, 0, and 1 for the first factor and 0, 1, 1 for the second factor. Also show that deriving predictions with discrete information is not a problem for the DPM.

8.3 Show that multiplicative and additive integration make very different predictions by assuming an additive integration rule in the FLMP. Work out the predictions of an "additive FLMP" using the parameter values given in problem 8.1. How do they differ from the multiplicative FLMP?

8.4 Using the research on letter recognition and reading described in this chapter as an example, develop an experiment that tests the DPM and FLMP in a different research domain. Suggestions for domains include object recognition, speech perception, and person impression.

8.5 In some respects, template models have always received a bum rap in psychology. As I tell my students, what are features? Nothing but mini-templates. In practice, template models are useful only for commentators who manipulate parts of things and expect to predict judgments of the whole based on the parts. How might template models be made to predict different results for different subjects?

8.6 Indicate why $P(G) = 0$ according to equation 6.3, when gap size is detected at not closed and line angle is detected as horizontal.

References

Fraser, J. (1986). Chaos: Solving the unsolvable, predicting the unpredictable. In M. F. Barnsley and S. G. Demko (Eds.), *Chaotic dynamics and fractals*. New York: Academic Press.

Gibson, E. J. (1969). *Principles of perceptual learning and development*. New York: Appleton.

Gleick, J. (1987). *Chaos: Making a new science*. NY: Penguin.

Green, D. J., and Fusella, J. M. (1972). A decompositional analysis of simple addition. *Psychological Review*, 79, 329–343.

Huey, E. B. (1908). *The psychology and pedagogy of reading*. 1968. Reprint. Cambridge, MA: MIT Press.

Journal of Experimental Psychology: Human Perception and Performance. (1994). Vol. 20, no. 6. Modeling visual word recognition.

Massaro, D. W. (1979). Letter information and orthographic context in word perception. *Journal of Experimental Psychology: Human Perception and Performance* 5, 595–609.

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

Massaro, D. W., and Cohen, M. M. (1991). Integration versus interactive activation: The joint influence of stimulus and context in perception. *Cognitive Psychology* 23, 558–610.

Massaro, D. W., and Cohen M. M. (1993). The paradigm and the fuzzy logical model of perception are alive and well. *Journal of Experimental Psychology: General* 122, 115–124.

Massaro, D. W., and Cowan, N. (1993). Information-processing models: Microscopes of the mind. *Annual Review of Psychology* 44, 383–425.

Massaro, D. W., and Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review* 97, 225–252.

Massaro, D. W., and Hary, J. M. (1986). Addressing issues in letter recognition. *Psychological Research* 48, 123–134.

McClelland, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology 23*, 1–44.

McClelland, J. L. and Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review 88*, 375–407.

McClelland, J. L., and Rumelhart, D. E. (1988). *Explorations in parallel distributed processing*. Cambridge, MA: MIT Press.

Oden, G. C. (1977). A fuzzy logical model of letter identification. *Journal of Experimental Psychology: Human Perception and Performance 3*, 336–352.

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

Sacks, O. (1970). *The man who mistook his wife for a hat and other clinical tales*. New York: Harper and Row.

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

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology 18*, 643–662.

Tulving, E., Mesulam, G., and Bancroft, G. (1966). Interaction of two sources of information in tachistoscopic word recognition. *Canadian Journal of Psychology 18*, 62–71.

Willows, D., Kruk, R., and Corcos, E. (Eds. 1993). *Visual processes in reading and reading disorders*. Hillsdale, NJ: Erlbaum.

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

About the Author

Dominic Massaro is chair of the Department of Psychology at the University of California, Santa Cruz. He was an undergraduate at UCLA and did his graduate work in psychology at the University of Massachusetts, Amherst.

Massaro's research interests include perception, memory, cognition, learning, and decision making. His current research is on the development and theoretical and applied use of a synthetic talking head for speech synthesis. He is also a passionate cyclist, and says that the beautiful Mount Tamalpais and the UCSC campus provide a persistent challenge and distraction for him.

> The process of scientific discovery turns me on. I get very excited when I think about a novel way of hitting on a breakthrough at how we and other living systems work. I love it when our commonly accepted beliefs are challenged and overturned by evidence. It is a real high when the results come pouring in revealing whether things worked the way you guessed. The competitive spirit is aroused when the results contain some surprises (or when your colleagues see things differently than you did). It is also highly rewarding to see students get turned on to this exciting field. Finally, the few times your ideas are actually implemented in practice represent an all-time high.
>
> In writing this chapter, I really enjoyed the challenge of making my work accessible to students and nonspecialists in the field. The process gave me an appreciation of the difficulties in communication in the world of science as well as in other worlds. One ironic benefit of this profession is that we get paid for these joyous travailes.