In D. Scarborough & S. Sternberg (Eds.) *An Invitation to Cognitive Science, Volume 4: Methods, Models, and Conceptual Issues.* Cambridge, MA : M.I.T. Press, 1998. Pp. 365-454.

Chapter 9

Inferring Mental Operations from Reaction-Time Data: How We Compare Objects

Saul Sternberg

Editors' Introduction

It is said of some people that they cannot walk and chew gum at the same time. To ask whether a person can do two things at the same time is an informal way of asking whether two or more processes or operations can be carried out in parallel (at the same time) or whether they must be done serially (one after another).

To begin to answer a question like this we need to focus on a specific task. Gum chewing and walking might be done in parallel, but not driving a car and reading a newspaper, although you may have seen some people try! Also we need to explore the task under conditions where we can control the situation and measure a person's behavior precisely. This chapter is a case study of one particular task: how people decide whether two objects are the same or different when they may be the same or may differ in one or more ways, for example, in size, color, or shape.

By manipulating how the objects to be compared differ, and by measuring how quickly observers can make accurate judgments in this task, Saul Sternberg shows how we can develop and test detailed theories of the mental operations that people carry out in this situation. His conclusion is perhaps surprising: although people make "different" judgments on the basis of serial processing, they make "same" judgments on the basis of some other sort of processing that is not yet fully understood. But because an observer never knows in advance whether a pair of objects will be the same or different, both kinds of processing may have to be employed for each judgment. That this conclusion does not seem intuitively obvious illustrates the importance of carefully testing theoretical ideas. Arriving at conclusions like this requires us to spell out in detail the properties of the mental operations assumed in each alternative theory.

Chapter Contents

The study of the time relations of mental phenomena is important from several points of view: it serves as an index of mental complexity, giving the sanction of objective demonstration to the results of subjective observation; it indicates a mode of analysis of the simpler mental acts, as well as the relation of these laboratory products to the processes of daily life; it demonstrates the close interrelation of psychological with physiological facts, an analysis of the former being indispensable to the right comprehension of the latter; it suggests means of lightening and shortening mental operations, and thus offers a mode of improving educational methods; and it promises in various directions to deepen and widen our knowledge of those processes by the complication and elaboration of which our mental life is so wonderfully built up.
—Joseph Jastrow in *The Time-Relations of Mental Phenomena* (1890)

9.1 Introduction

You frequently have to decide whether something is a particular object. (Is the person at the door the friend you expected? Is that car yours? Is this the book you had planned to take? Is that a stop sign?) Your decision in such cases is usually both rapid and accurate. How do people accomplish this feat?

The object-comparison experiment is one of the simplest used to study the perception of visual patterns. On each of a series of trials, two clearly visible patterns are presented successively, or, in some experiments, simultaneously in adjacent positions. The subject in the experiment must make one of two responses: If the patterns are identical (relative to a specified set of criteria), the "same" response, R_{same}, is correct; if the patterns differ, the "different" response, R_{diff}, is correct. The responses might be spoken words ("yes" or "no," for example) or button presses (left button or right button, for example). Pattern differences are sufficiently great when they occur, and the patterns sufficiently simple, that without time pressure all responses are likely to be correct. However, the subject is instructed to respond as rapidly as possible consistent with high accuracy (and may be rewarded for doing so). The time from the onset of the display of the second pattern to the subject's response is the "reaction time" or RT; in a typical experiment the mean (average) reaction time, $\overline{RT}$[1], might range from 400 to 500 ms, depending on the pair of patterns; because of the

time pressure a subject might make errors on a small proportion (perhaps 3 percent) of the trials. We make inferences about the underlying processes of pattern perception and comparison by examining the ways in which manipulations of the pair of patterns influence the *RT*.

Why measure reaction times? I am tempted to reply, "Because they are there," and leave it at that, but this is not quite my opinion. I believe that the time measurements that will best reveal aspects of mental functioning are collected, not by observing reaction times that "are there" in everyday life, but in carefully constructed laboratory situations where conditions can be well controlled and subjects can be motivated as described above. The tasks studied under these conditions are often ones that seem automatic, effortless, and instantaneous to the person performing them, and can be done with high accuracy. Nonetheless, the times they take are substantially longer than the neural transmission times for input (eye to brain) and output (brain to response), and vary systematically with manipulations of the task. Examples of such tasks are naming a visually presented word, deciding whether a word like *chair* is the name of an animal, searching a small array of letters for a target letter, deciding whether a particular number is contained in a previously memorized list, producing a memorized utterance, or, as in the present chapter, deciding whether one pattern is the same as another. Although these tasks are relatively simple, researchers believe they reveal capacities and limitations that underlie mental activities in everyday life.

Thus, in the approach to the analysis of cognitive processes considered in the present chapter, we examine these processes under conditions where they function virtually without error. By applying time pressure to the subject under these conditions, the experimenter hopes to induce some of the mechanisms at work to reveal themselves, not by how often or in which ways they fail—an alternative approach to the analysis of mental processes—but by how much time they need in order to succeed.

What is special about the object-comparison experiment? There are several appealing reasons for considering experiments of this type in an introductory chapter on the use of reaction-time data. A remarkably powerful set of inferences can be made from some straightforward and simple aspects of the data from such experiments, when stimuli are used that have been explicitly constructed to have several attributes (such as the size, shape, and color of a geometric form) or several elements (such as the letters in a letter string). Analysis shows how simple but interesting theories yield predictions that can then be tested by examining reaction-time patterns. And at the same time as some aspects of the data are fairly decisive in selecting among alternative theories, other aspects reveal interesting puzzles that have yet to be convincingly resolved.

The object-comparison experiment is not only interesting in itself, but also provides an excellent vehicle for learning about some of the impor-

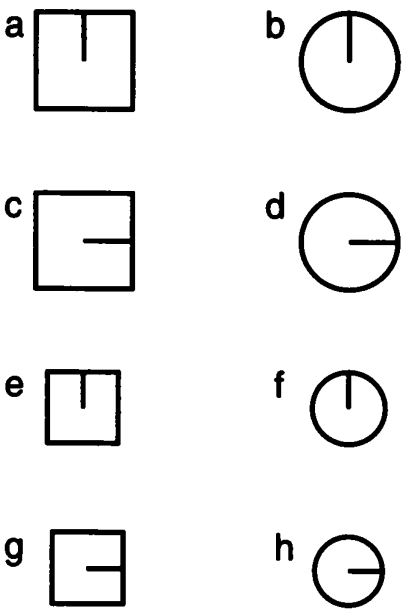

Figure 9.1
Set of eight simple geometric patterns that might be used in multiattribute object comparison experiment. The patterns vary in area, shape, and direction of the included line.

tant issues that arise in working with reaction-time data, for becoming familiar with and developing intuitions about serial and parallel processing mechanisms, for appreciating how theories of such mechanisms must be elaborated to make them testable, and, in general, for practice in making inferences from behavioral data to underlying mental mechanisms. These goals have led me to concentrate almost exclusively on the results from just one beautiful experiment; I mention other data only to illuminate some of the important issues that are not addressed by that experiment and to indicate the generality of the phenomena it reveals.

9.1.1 A Three-Attribute Stimulus Set

The typical patterns used in object-comparison experiments are simple compared to most everyday objects, but researchers believe that they capture an important property of such objects: they vary on several "attributes." For example, figure 9.1 shows eight simple geometric patterns that differ in area (large, small), shape (square, circle), and direction of the line segment (up, right). The attributes are thus area (A), shape (S), and direction (D); and, in this example, each attribute can have one of two "values."[2] The two values that the attribute shape can take on in this case, for example, are "square" and "circle." Furthermore, because the attributes vary independently, there are eight (2^3) possible combinations. I will usually refer to the values of attributes, such as squareness, as "features" of the stimulus.

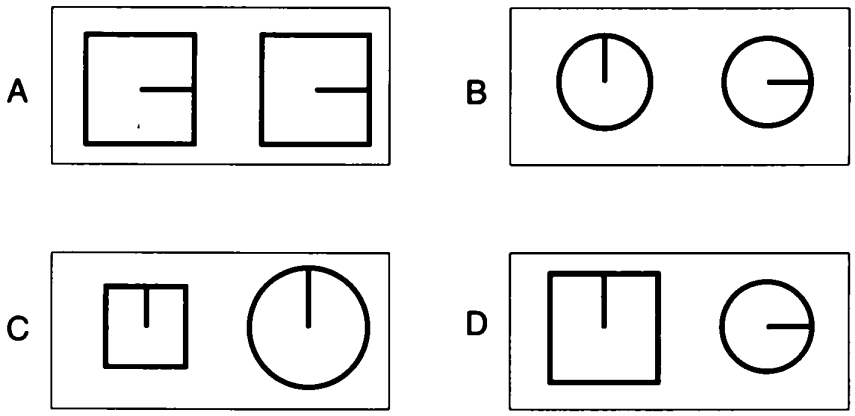

Figure 9.2
Four examples of pairs of patterns that might be presented on a trial. For pairs A, B, C, and D, $n_{diff} = 0$, 1, 2, and 3, respectively.

In a condition in the object-comparison experiment in which all three attributes are "relevant," ($n_{rel} = 3$), the subject would be asked to respond "same" if the values of all three attributes are the same (i.e., if $n_{diff} = 0$) and to respond "different" otherwise (i.e., if $n_{diff} \geq 1$). Figure 9.2 shows four examples of pairs of patterns. Response R_{same} is correct for A, which includes pattern (c) from figure 9.1, repeated; $n_{diff} = 0$. Response R_{diff} is correct for B, which includes (f) and (h) from figure 9.1 (one attribute—direction—differs; $n_{diff} = 1$), for C, which includes (e) and (b) from figure 9.1 (two attributes—area and shape—differ; $n_{diff} = 2$), and for D, which includes (a) and (h) from figure 9.1 (all three attributes differ; $n_{diff} = 3$).

In another typical condition, only two of the attributes would be relevant, $n_{rel} = 2$. This could be achieved, with the relevant attributes being area and shape, by holding direction constant, restricting the set of patterns, for example, to the four shown in figure 9.1 with direction right: (c), (d), (g), and (h). The subject would be informed that this restriction was in force, and provided practice in the restricted task. Thus, if the pair (c) and (c) were presented, the subject would not have to compare the directions of the two patterns to determine that they were the same. A second way to create a condition with the two relevant attributes area and shape would be to permit direction to vary, and use the full set of eight patterns shown in figure 9.1, but to instruct the subject to ignore direction in determining "same" and "different." Any difference in direction between the two patterns then becomes irrelevant. In this experimental condition, then, R_{same} would be correct for the pair (c) and (c), as above, but also for the pair (c) and (a), for example. Because n_{diff} is the number of features that differ only among those that are relevant, $n_{diff} = 0$ in both of these cases.

A second type of pattern used in object-comparison studies consists of a string of letters, and is probably better described as a set of elements

with identities, rather than a set of attributes with values. To simplify the discussion, I shall use "attribute" and "feature test" in an extended sense, also to mean element and identity test.

9.1.2 Major Issues in Comparing Multiattribute Objects

Starting with Egeth's pioneering study (1966), the primary question in experiments like these is how information from several attributes is used in the discrimination of visual objects. One way to put the question is similar to the way Egeth (p. 245) did: Do humans "discriminate between multi-attribute objects by comparing them one attribute after the other (sequential mode), or by comparing them on several attributes simultaneously (parallel mode), or by comparing unitary representations of them without regard to their component attributes (template mode)"? Or is none of these simple possibilities correct?

9.1.2.1 Holistic versus Feature Comparison

Another way to phrase Egeth's question is to ask two related questions. First, is the pattern comparison process "holistic" (patterns compared as wholes) or is it "analytic" (Nickerson 1972), depending, instead, on analysis into separate features (e.g., direction: up, and area: large), together with separate tests of those features? (I shall call the process by which it is determined whether values on corresponding attributes match or mismatch a "feature test" or just a "test.")

9.1.2.2 Sequential versus Parallel Tests

Second, if feature tests are involved, are they carried out sequentially, or in parallel? One important development of the last two decades in the understanding of the brain is the view that the visual system separates patterns into their constituent features, which are coded and processed by distinct mechanisms in different parts of the brain.[3] This adds plausibility to the idea that pattern comparison might depend on feature analysis; it also raises the question whether a given set of stimulus attributes or features is psychologically and neurologically "real"—that is, corresponds to aspects of patterns that are in fact processed separately.

Sequential and parallel theories lead to predictions that can be tested with data from the object-comparison experiment, predictions that answer four important questions.

1. How does the time taken to decide "same," $\overline{RT}_{same}$, vary with the number of relevant attributes, n_{rel}?
2. How does the time taken to decide "different," $\overline{RT}_{diff}$ for a given n_{rel}, vary with the number of those n_{rel} attributes on which values differ, n_{diff}?

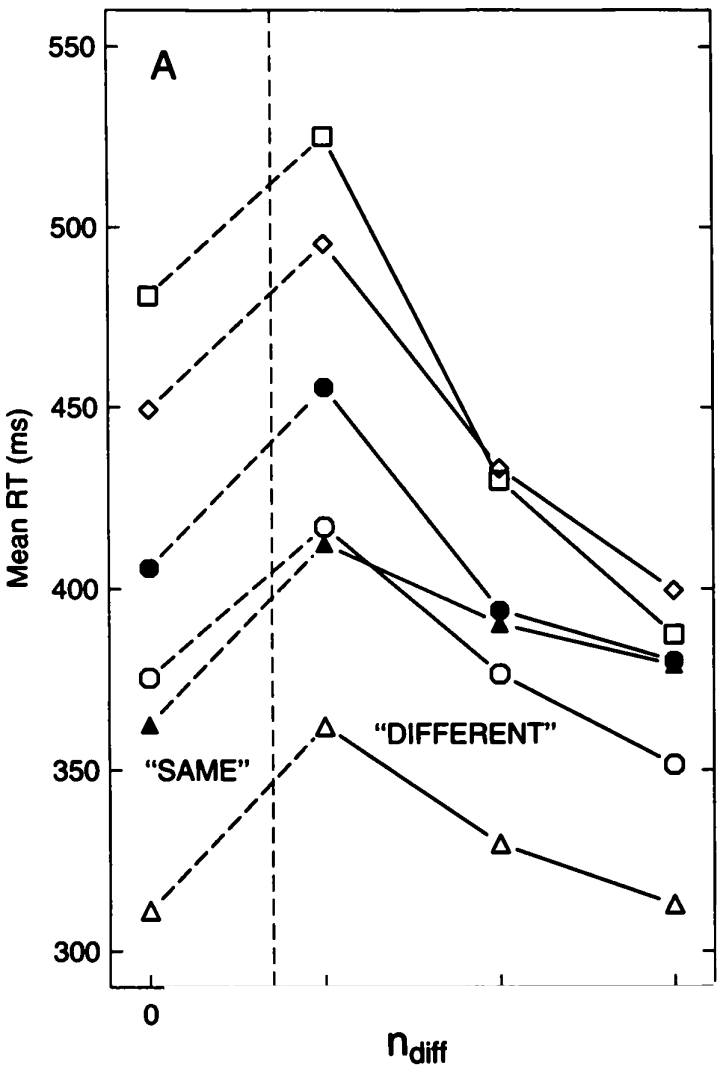

Figure 9.3
Mean data reviewed by Nickerson (1972, figure 10) from six conditions in three object-comparison studies using simple geometric patterns. In all conditions, three stimulus attributes were relevant ($n_{rel} = 3$), and same and different pattern pairs were equally frequent. Details of the experiments differed: for example, different attributes were used, and the discriminations differed in difficulty. Selecting attributes that differed markedly in difficulty, Hawkins (1969, experiment 1) used a "go/no-go" procedure, such that in some sessions subjects either produced R_{same} or made no response, while in other sessions they either produced R_{diff} or made no response. Nickerson's experiment (1967) reported data for two levels of practice and included conditions in which the two patterns were presented serially (SER), and other conditions in which they were presented simultaneously (SIM), as Egeth (1966) and Hawkins (1969) presented them. Panel A shows $\overline{RT}$ for correct responses; panel B shows the percentage of trials on which an error occurred. (Some error data from Hawkins's experiment and all error data from Egeth's are missing.) Both are plotted against the number of attributes along which the pair of stimuli differed (n_{diff}). When this number is zero, R_{same} is correct; when this number is greater than zero, R_{diff} is correct; responses plotted to the left and right of the vertical broken line therefore differ.

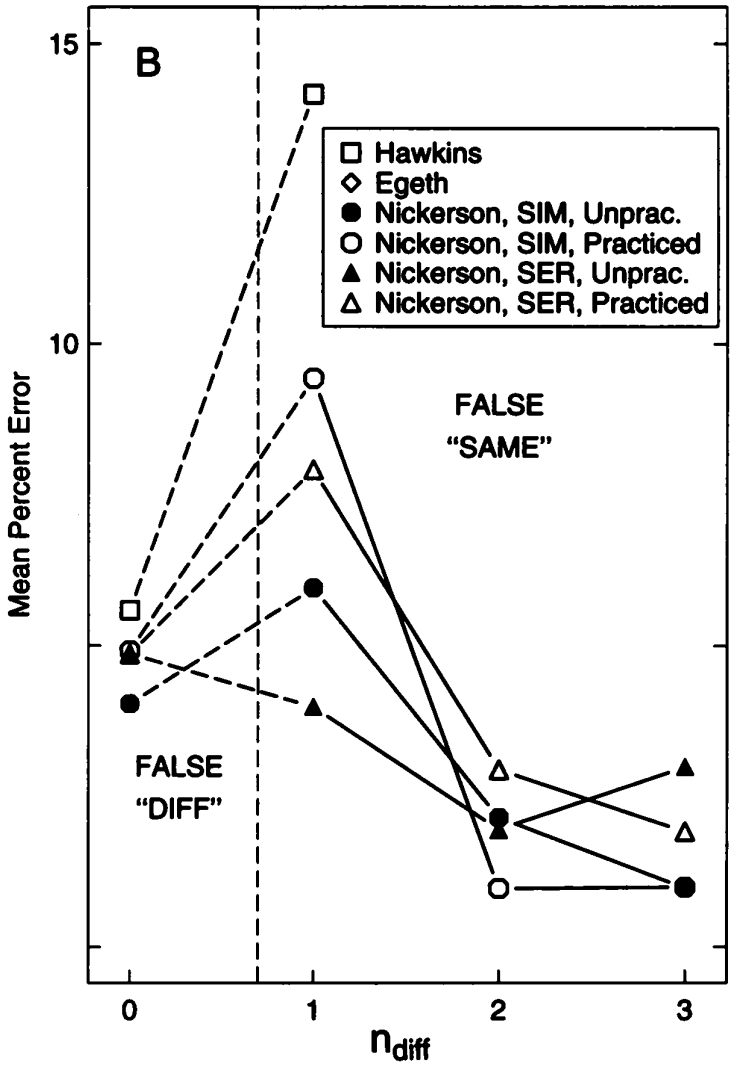

Figure 9.3 (cont.)

3. How does the time to decide "different," $\overline{RT}_{diff}$ for a given n_{diff}, vary with the number of relevant attributes, n_{rel}?
4. What is the relation between the $\overline{RT}$s for R_{same} and R_{diff}?

9.1.3 Some Typical Data

9.1.3.1 Data from Geometric Patterns

Six sets of data from some of the studies reviewed by Nickerson (1972) are shown in figure 9.3. Considering the procedural differences among the conditions and the substantial differences in overall $\overline{RT}$, the similarity among the six sets of RT data is striking. Although the pattern of errors is also of interest, and may help to select among alternative theories, this

chapter outlines only attempts to explain the effects of n_{diff} and n_{rel} on $\overline{RT}$; the RTs considered are just those for the correct responses.[4]

9.1.3.2 Data from Letter Strings

Three years after Egeth's research appeared, Bamber (1969) generalized and extended Egeth's findings. Instead of simple geometric patterns, Bamber used letter strings, of length 1, 2, 3, or 4, with the two strings on a trial presented successively.[5] Suppose the length is 3. Then R_{same} would be correct for the pair **KSV**, **KSV**, while R_{diff} would be correct for the pair **KSV**, **KTV**.[6] Here the object is a letter string, the element is a letter in a position, and the identity of the letter plays the role of a feature in a geometric pattern. The number of relevant elements, n_{rel}, is the number of letters in the string.

Bamber's results are shown in figure 9.4. Unlike the data in figure 9.3, where $n_{rel} = 3$, they include variation in n_{rel}. Why is it interesting to compare Bamber's letter-string experiment, where the stimulus component manipulated in the experiment is a letter within a string of distinct letters, with Egeth's pattern experiment, where it is an attribute such as the size or shape of a single object? There are at least two important differences between these two sorts of stimuli—the ease of analyzing them into their components and the variation in their complexity. Subjects have had a great deal of experience with each letter in a large range of contexts before entering the laboratory. Together with the separateness of letters within a string, this makes it plausible that letter strings might be readily parsed into their component letters, and the letters processed separately. Evidence suggests that attributes such as the size or shape of a single object need not be treated in this way. If the patterns of data are similar (as they are), this would support the hypothesis that the attributes that are varied in geometric pattern experiments are indeed analyzed separately.

With respect to complexity, an increase in the number of relevant attributes, n_{rel}, in a pattern experiment need not be associated with an increase in stimulus complexity, especially if the total number of attributes is held constant, as in figure 9.1. On the other hand, any reasonable measure of stimulus complexity would say that it increases with the number of elements in a letter-string experiment; this could introduce difficulties into the interpretation of letter-string experiments. Again, if the data from the two kinds of experiment are similar, this perhaps suggests that such complexity variations are contributing relatively little to the data pattern.

9.1.4 Plan of the Chapter

The data described above have two remarkable aspects. First, the ways in which $\overline{RT}_{diff}$ varies with n_{rel} and n_{diff} can be nicely explained by a rela-

tively simple theory. Second, there seems to be no single simple theory that can also explain the way in which $\overline{RT}_{same}$ varies with n_{rel}. These two aspects of what we know have dictated the organization of the present chapter. I start with an attempt to explain the R_{diff} data (section 9.2), referring to R_{same} only where development of the sequential-test theory for R_{diff} makes it expedient. I then turn to a systematic discussion of the R_{same} data (section 9.3).

For the R_{diff} data, I explore a theory of sequential tests, which turns out to work well. The exploration has two parts. First, sections 9.2.1–9.2.2 are concerned with what the theory says about the average number of feature (or element) tests, $\bar{n}_{tests}$, required for an accurate decision. I have included several tables that contain numerical examples of the important phenomena generated by the theory, to aid your comprehension and educate your intuition. Second, section 9.2.3 is concerned with what can be said about the behavior of $\overline{RT}$, given the behavior of $\bar{n}_{tests}$. The leap from $\bar{n}_{tests}$ to quantitative statements about $\overline{RT}$ turns out to be surprisingly subtle, and requires us to consider several particular variants of the sequential-test theory, each imposing different constraints on the mechanism. With none of these constraints the theory can predict very little.

To check whether our evaluation of the theory has been sufficiently probing, we need to make sure that the same data cannot also be explained by a fundamentally different competing theory. The alternative considered in sections 9.2.5–9.2.7 is a theory of parallel tests, one where the tests are *independent*, having no influence on each other. While some aspects of the data can be equally well explained by this very different theory, others cannot; this forces us to augment the parallel-test mechanism with a sequential encoding process.

Turning to the $\overline{RT}_{same}$ data, I show in section 9.3 that the simple sequential-test theory that accounts quite well for the $\overline{RT}_{diff}$ data cannot handle them. We are forced to consider the possibility, strange as it may seem, that two different processes (P_{same} and P_{diff}) give rise to the two responses (section 9.4). Assuming two processes, the data indicate how they are arranged in time. Also, it can be argued that P_{same} is neither of the analytic mechanisms we have considered—sequential or parallel tests. The chapter ends with some discussion of the possible nature of P_{same}.

The first of two appendices is concerned with the role of errors during reaction-time experiments in the interpretation of the RT data. The second appendix relates what we have been doing to the classical *subtraction method* for the analysis of mental processes. Because it is convenient to introduce several symbols along the way, I have included a glossary.

> *Comment 1: Role of comments.* I find the footnotes in technical material to be among the most interesting parts because they often say

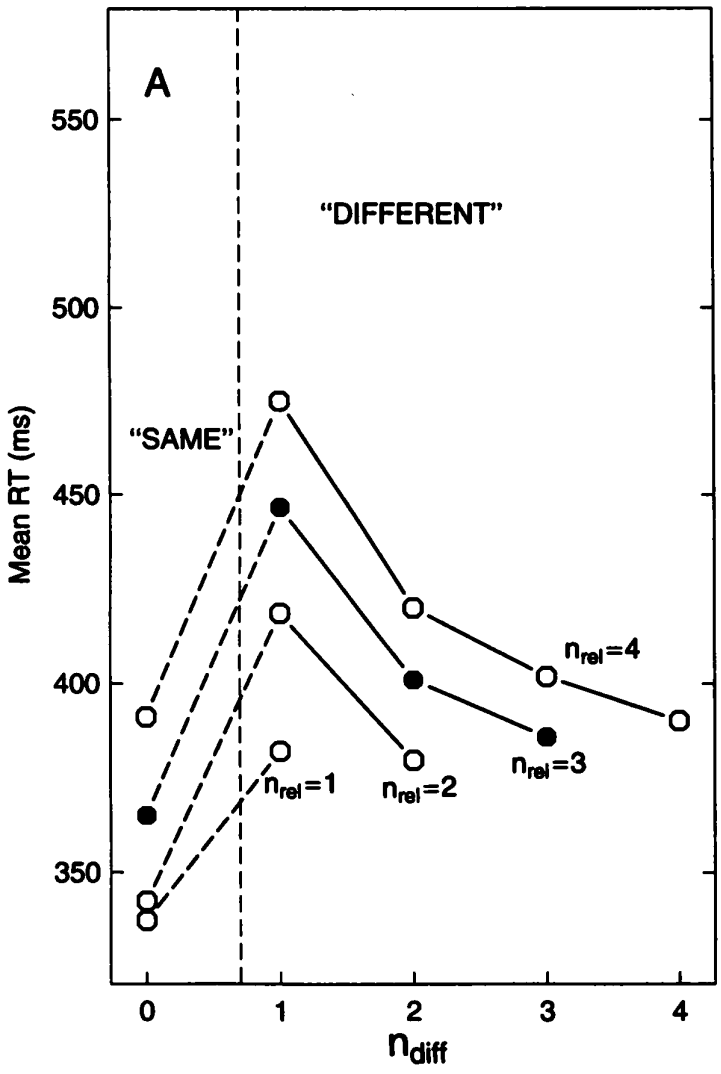

Figure 9.4
Mean data from Bamber's object-comparison experiment (1969), in which patterns were strings of letters. On "different" trials, different numbers of letters were displayed and all were relevant: $1 \leq n_{rel} \leq 4$. Panel A shows $\overline{RT}$ for correct responses; panel B shows the percentage of trials on which an error occurred. Both are plotted against n_{diff}, the number of letters in corresponding positions that differed ($0 \leq n_{diff} \leq n_{rel}$). When $n_{diff} = 0$, R_{same} is correct; when $n_{diff} \geq 1$, R_{diff} is correct. Each of four subjects contributed data from about 2,800 trials; each data point in the figure is based on 480 or more trials. Points for $n_{rel} = 3$ are distinguished to aid comparison with figure 9.3.

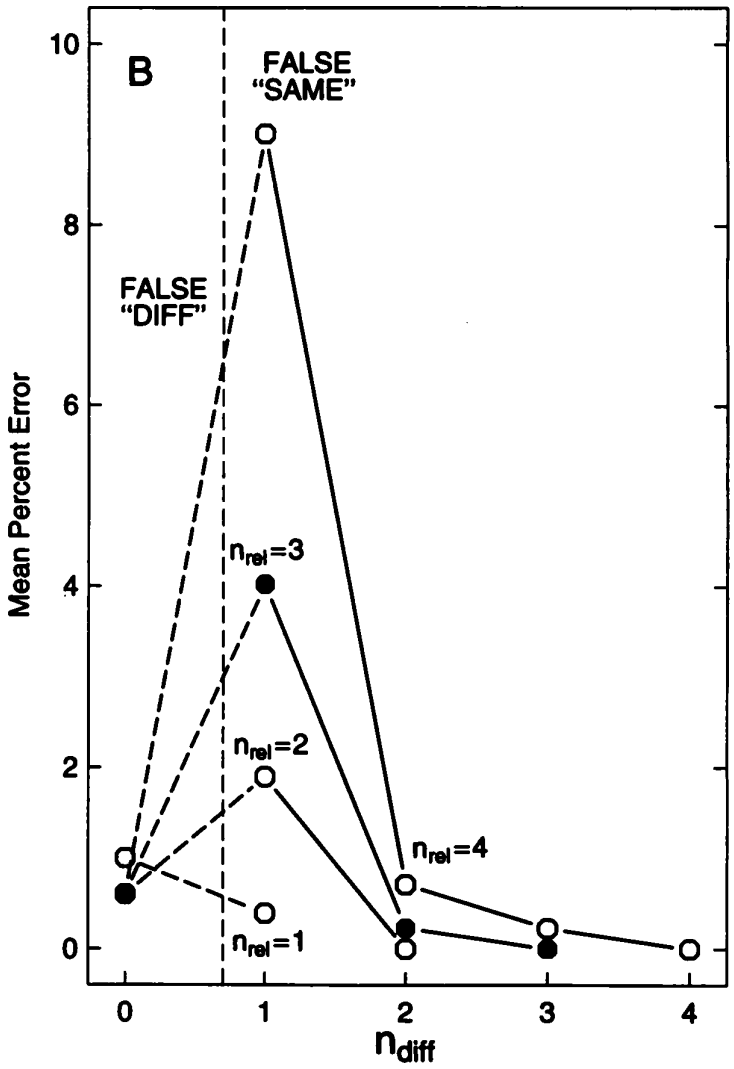

Figure 9.4 (cont.)

more about the author's perspective on the subject than does the text itself. Also, like hypertext, footnotes acknowledge the multiplicity of levels to be communicated, which are not easily expressed in linear narrative. Especially given recent literary trends, where footnotes appear even in fiction (e.g., Baker 1988) and where novels contain abrupt alternations of viewpoint (e.g., Calvino 1981), it seems desirable in technical material to be explicit about multiple levels. Hence, to step away from the text and provide some perspective, without necessarily imposing undersized type on the reader, I occasionally make use of "comments" set as extracts in the running text of this chapter (and chapter 14). Skipping over a comment should not interfere with comprehension of the text that follows, and is certainly reasonable on a first reading.

This chapter is a long one; readers who do not fully share my fascination with attempts to understand the object-comparison experiment are encouraged to read selectively. One possibility would be to read the introduction (section 9.1), the basics of sequential testing (sections 9.2.1–9.2.3, 9.2.4.1, 9.2.4.7), an introduction to parallel testing (sections 9.2.5, 9.2.6), and a comparison of the two theories as applied to "different" responses (section 9.2.9). The selective reader could then read about the difficulties presented by "same" responses for sequential testing (section 9.3.1) and about an approach toward their resolution (section 9.4.1).

9.1.5 Theories, Models, and Data

Before proceeding, let us take a brief detour to consider the enterprise as a whole. Each theory of visual-comparison performance we consider asserts that a particular mental process or mechanism is responsible for making the required same-different decision. Usually we start with a vague idea about a possible mechanism: for example, just the idea that the corresponding features or elements of the two objects are compared sequentially. We often then discover that to render the theory testable, the mechanism has to be specified more precisely, elaborating it beyond the initial vague idea. The more precise description that results is sometimes called a "model" of the mechanism. Each of the mechanisms we consider consists of a set of defining properties, together with a set of derived properties ("predictions") that follow from them. The defining properties of models are sometimes called "assumptions"; this is not intended to protect them from being questioned and tested. A theory or model *explains* a set of data to the extent that model properties and data properties agree. We wish to find hypothetical mechanisms that are plausible and simple, yet capable of predicting a relatively large number of derived properties we can then compare with data. Such comparisons of properties of the hypothetical mechanism with properties of the data have two functions: to help decide how well the model is likely to approximate the (real) mechanism that underlies the data; and to consider the model as a "baseline," using the structure of the deviations of data properties from model properties to gain further understanding of the actual mechanism.

As the examples in this chapter will show, there are often aspects of a set of data that can be explained equally well by theories that are conceptually very different. For example, we will see that both of the major theories we consider for "different" responses can predict that $\overline{RT}_{\text{diff}}$ varies directly with n_{rel} and inversely with n_{diff}. One helpful strategy to reduce the chances of being deceived by the success of a theory is to test quantitative predictions rather than qualitative ones wherever possible: a theory may predict a given direction for the effect on $\overline{RT}$ of some experi-

mental variation, as for the examples just mentioned, but is less likely to predict a particular quantitative form for the effect—for example, that $\overline{RT}_{diff}$ will decrease with a particular amount of deceleration as n_{diff} increases. A second helpful strategy is to put two or more theories into explicit competition; this forces the researcher to discover *discriminating properties* of the data—properties that only one of the competing theories can explain. This second strategy, of considering more than one theory concurrently, also reduces the risk of bias in the investigators (us). It is because of the importance of doing so that I have chosen to juxtapose alternative contrasting theories relatively early in the chapter, even before a systematic discussion of the R_{same} data.

> *Comment 2: Method of multiple working hypotheses.* The second strategy exemplifies Chamberlin's "method of multiple working hypotheses" (1890) to reduce the biasing influence of "parental affection" for a hypothesis. Chamberlin, a geologist, described several consequences of such affection, which, he argued, are less likely to arise if we maintain concurrent interest in more than one theory for a set of phenomena: "There is an unconscious selection and magnifying of the phenomena that fall into harmony with the theory and support it, and an unconscious neglect of those that fail of coincidence. The mind lingers with pleasure upon the facts that fall happily into the embrace of the theory, and feels a natural coldness toward those that seem refractory.... There springs up, also, an unconscious pressing of the theory to make it fit the facts, and a pressing of the facts to make them fit the theory" (755).

9.2 Reaction Time to Judge "Different"

We consider first the pattern of mean reaction times for R_{diff}. Figures 9.3A and 9.4A (right-hand portions) show that, given a fixed number n_{rel} of *relevant features* (features that might or might not differ), $\overline{RT}_{diff}$ decreases as the number n_{diff} of mismatching features increases. This is the first phenomenon to be explained. The second phenomenon is not shown in figure 9.3, but has also arisen in experiments with multifeature patterns such as those in figure 9.1, and is shown in figure 9.4A: for a fixed number n_{diff} of differing features, $\overline{RT}_{diff}$ increases as the number n_{rel} of relevant features increases. How can we explain these two phenomena?

9.2.1 Sequential Tests: Defining Properties

We can think of a sequence of feature tests as a search, where the *target* of the search is "any mismatch." The search is "self-terminating" if, in the

event of a mismatch (i.e., the presence of a target), the response is initiated as soon as the target is found, rather than awaiting completion of all the tests. We shall see that such self-terminating sequential feature (or element) testing can account for the two phenomena mentioned above with quantitative precision.

For sequential tests let us assume the following for each trial:

1. Tests are carried out one at a time;
2. The mean duration of a test is unaffected by the number of other tests that must be carried out;
3. No test is carried out more than once;
4. R_{diff} is initiated if and as soon as a feature mismatch is discovered (the process is self-terminating); or
5. R_{same} is initiated if and as soon as all n_{rel} tests are completed with no mismatch.

On any trial, the search for a mismatch proceeds through the features in some order, an order I shall call the "search path." In deriving predictions for the sequential model, a critical additional property is often added:

6. Random-order property: If there are any mismatching features, they are located at random positions in the search path.

That is, all the possible positionings of the mismatching features are equally likely. For example, if $n_{\text{rel}} = 3$ and $n_{\text{diff}} = 1$, then the mismatching feature is assumed to be located equally often in the first, second, and third positions in the search path. Fortunately, it is possible to design the experiment so as to satisfy this assumption, regardless of the subject's choice of search path from trial to trial (which we might well be unable to infer). Suppose three features are varied in the experiment, for example, and consider trials on which the patterns in the pair differ by one feature. The trial sequence can be constructed so that mismatches occur, unpredictably, for each of the three features. If the data are averaged, with equal numbers of these three kinds of trial, then the requirement that the mismatch be located at a random position in the search path will be satisfied, whatever search order or mixture of search orders the subject adopts. We must, of course, also assume that the search path on a trial is not influenced by which features match and which features mismatch on that trial.

We next consider the derived properties of the sequential mechanism, in two steps: First, how does the *number of tests* required for a "different" response, $n_{\text{tests}}(\text{diff})$, depend on the numbers of relevant and mismatching features, n_{rel} and n_{diff} (section 9.2.2)? Second, how does our measure, $\overline{RT}$, depend on the predicted number of tests (section 9.2.3)?

In examining a model, some properties (here the "defining properties") are regarded as essential. As we shall see, auxiliary properties (called

"constraints" in section 9.2.3) often have to be added to extract predictions. One of the challenges in evaluating a model is to assess the relative importance of the defining versus auxiliary properties in determining its success or failure.

9.2.2 Sequential Tests: Prediction of the Number of Tests

9.2.2.1 Effect of Number of Mismatching Features on Number of Tests

Let $n_{\text{tests}}(\text{diff})$ be the number of tests on a "different" trial up to and including the first mismatch. According to the model, the last of the sequence of tests associated with a "different" response (R_{diff}) is, of course, a mismatch; it is preceded by from 0 to $n_{\text{rel}} - 1$ matching tests, denoted n_{mtests}. Thus

$$n_{\text{tests}}(\text{diff}) = n_{\text{mtests}}(\text{diff}) + 1, \qquad (n_{\text{tests}} \leq n_{\text{rel}}). \tag{9.1}$$

The decline in $\overline{RT}_{\text{diff}}$ as n_{diff} increases may be understood in terms of the model by noting that on average, the more mismatching features, the earlier in the search path the first of them will be encountered. Thus the first of $j + 1$ mismatches among n_{rel} relevant attributes will be encountered earlier in the search path (on average) than the first of j mismatches. Suppose three attributes were relevant, $n_{\text{rel}} = 3$, for example, area (A), shape (S), and direction (D). In table 9.1 are listed the seven possibilities on a trial on which the features are tested in the order A → S → D. The first column shows the number of mismatching features, and determines the division of the table into three parts. The second column lists the particular features that mismatch. For example, given that $n_{\text{diff}} = 2$, the possible

Table 9.1
Effect of number of mismatching features n_{diff} ($1 \leq n_{\text{diff}} \leq 3$) on mean number of tests underlying "different" responses when $n_{\text{rel}} = 3$ and the search path is A → S → D. Matches are shown in lowercase italics, mismatches in uppercase bold.

n_{diff}	Mismatching feature(s)	Tests	$n_{\text{tests}}(\text{diff})$	$\bar{n}_{\text{tests}}(\text{diff})$
	A	**A**	1	
1	S	*a*,**S**	2	2.00
	D	*a*,*s*,**D**	3	
	A,S	**A**	1	
2	A,D	**A**	1	1.33
	S,D	*a*,**S**	2	
3	A,S,D	**A**	1	1.00

pairs of features that might mismatch are A-S, or A-D, or S-D. The third column shows the sequence of tests that occur up to and including the first mismatch. Thus, with S-D mismatching, the first test (of A) finds a match (the "*a*" shown in lowercase italics), but the second test (of S) finds a mismatch (the "**S**" shown in uppercase bold), which leads to the initiation of R_{diff}. Thus $n_{\text{tests}} = 2$ in this case, but $n_{\text{tests}} = 1$ in the other two-mismatch cases (A-S and A-D), as shown in column 4. Finally, the mean of the entries 1, 1, and 2 in column 4 yields $\bar{n}_{\text{tests}} = 1.33$ in the last column.

In practice, of course, we usually do not know the order in which features are tested—the search path. As there are six possible orders of testing A, S, and D, there will be six tables like table 9.1, one per search path. However, the values of $\bar{n}_{\text{tests}}$ in each of these tables will be the same. For example, in the second section of every one of these tables (in which $n_{\text{diff}} = 2$), the fourth column will contain two 1s and one 2, and $\bar{n}_{\text{tests}}$ will be 1.33.

Table 9.1 shows that the mean number of tests required to achieve a mismatch decreases as the number of mismatching features increases. The decrease in $\bar{n}_{\text{tests}}$ from 2 to 1.33 to 1 is nonlinear: instead of having a constant effect, an increase by one in the number of mismatching features ($n_{\text{diff}} \rightarrow n_{\text{diff}} + 1$) has a diminishing effect as more features mismatch (i.e., with larger n_{diff}). To anticipate the connection between $\bar{n}_{\text{tests}}(\text{diff})$ and $\overline{RT}$, if we assume that the mean duration of each feature test does not change as the number of feature tests changes, then $\overline{RT}_{\text{diff}}$ would decrease in the same way (decelerating) as does $\bar{n}_{\text{tests}}$. Figures 9.3A and 9.4A show just such a decelerating pattern as n_{diff} increases from 1 to 3.

> *Comment 3: Invariant test durations.* The assumption that mean feature-test duration is unaffected by n_{tests} is actually quite strong; it requires two conditions to be satisfied: first, the duration of a particular test must not be influenced by the other tests that accompany it. This *pure-insertion assumption* is discussed further in section 9.2.3.1 and appendix 2. And second, either (1) the order in which attributes are tested is random or (2) the mean duration of a test is the same regardless of which attribute is tested. (Condition 1 would not be satisfied if, for example, shape tended to be tested before size. And condition 2 would not be satisfied if, for example, a shape test took longer than a size test.) Given condition 1, the first attribute tested is equally likely to be any one of the set of relevant attributes, and likewise for the second feature tested, and so on. It follows that the mean duration of the first test is the same as the mean duration of the second, and so on. Given the alternative, condition 2, the mean duration of the kth test, $k = 1, 2, \ldots, n_{\text{tests}}$, is not influenced by k.

Table 9.2
Effect of number of relevant features n_{rel} ($2 \leq n_{rel} \leq 4$) on mean number of tests underlying "different" responses when $n_{diff} = 2$

n_{rel}	Search path	Mismatching features	Tests	n_{tests}(diff)	$\bar{n}_{tests}$(diff)
2	A → S	A,S	**A**	1	1
3	A → S → D	A,S	**A**	1	1.33
		A,D	**A**	1	
		S,D	***a*,S**	2	
4	A → S → D → C	A,S	**A**	1	1.67
		A,D	**A**	1	
		A,C	**A**	1	
		S,D	***a*,S**	2	
		S,C	***a*,S**	2	
		D,C	***a*,*s*,D**	3	

Thus conditions 1 and 2, which are discussed further in section 9.2.4.5, separately imply that the mean test duration is independent of its position in the test sequence. And, given pure insertion together with either (1) or (2), it follows that mean test duration is unaffected by n_{tests}.

9.2.2.2 Effect of Number of Relevant Features on Number of Tests

What happens to the number of feature tests as n_{rel}, the number of features that might differ, and that therefore may have to be tested, increases? For illustration, let "C" represent a fourth attribute, color, and suppose the patterns have two mismatching features, $n_{diff} = 2$; consider what happens as the number of relevant features increases: $n_{rel} = 2, 3, 4$. The third column in table 9.2 lists all possible pairs of mismatching features for each of these three values of n_{rel}. For each n_{rel} a particular search path is assumed. Given the search path and the pair of mismatching features, the tests required are listed in the fourth column, and the number of these tests in the fifth. The mean number of tests, $\bar{n}_{tests}$, (with equal numbers of trials for each set of mismatching features), is shown in the last column for each value of n_{rel}; this mean number is the same for any search path. We thus see that the mean number of tests required to achieve a mismatch increases as the number of relevant features increases. As suggested by the values of $\bar{n}_{tests}$ in the last column, the increase is linear: an increase by one in the number of relevant features ($n_{rel} \rightarrow n_{rel} + 1$) increases $\bar{n}_{tests}$ by the same 0.33 for any number n_{rel} of relevant features.

9.2.2.3 *A General Statement of the Two Effects on Number of Tests for "Different" Responses*

The examples described above should provide an intuitive basis for the following general statement. For the "different" trials in a self-terminating sequential testing process with the mismatching features located at random points in the search path, $\bar{n}_{\text{tests}}(\text{diff})$ depends on n_{diff}, the number of mismatching features, and n_{rel}, the number of relevant features:[7]

$$\bar{n}_{\text{tests}}(\text{diff}) = \frac{n_{\text{rel}} + 1}{n_{\text{diff}} + 1}, \qquad (n_{\text{rel}} \geq n_{\text{diff}} \geq 1). \tag{9.2}$$

In equation 9.2, the denominator represents the influence of n_{diff} illustrated in table 9.1; $\bar{n}_{\text{tests}}$ decreases nonlinearly as n_{diff} increases. The numerator represents the influence of n_{rel} illustrated in table 9.2, in which $\bar{n}_{\text{tests}}$ increases linearly with n_{rel}. From equation 9.1, the mean number of *matching* tests is $\bar{n}_{\text{mtests}}(\text{diff}) = \bar{n}_{\text{tests}}(\text{diff}) - 1$. The points connected by solid lines in figure 9.5 show this quantity plotted as a function of n_{rel} for each of four values of n_{diff} ($1 \leq n_{\text{diff}} \leq 4$).

The figure illustrates four properties of $\bar{n}_{\text{tests}}$:

1. $\bar{n}_{\text{tests}}$ increases linearly with n_{rel}. (The points for $n_{\text{diff}} = 1$, 2, and 3 fall on straight lines with positive slopes);
2. $\bar{n}_{\text{tests}}(\text{diff})$ decreases nonlinearly and with "diminishing returns" as n_{diff} increases (for each value of n_{rel} the point for a larger value of n_{diff} lies below the point for a smaller value, and the separation between such points is decreasing);
3. If $n_{\text{diff}} = n_{\text{rel}}$ (the filled points in the figure), then, because the first test must be a mismatch, $n_{\text{tests}} = 1$ for all values of n_{rel};
4. The rate at which $\bar{n}_{\text{tests}}$ grows with n_{rel} is reduced as we increase n_{diff} (the slopes of the four lines, given by $1/(n_{\text{diff}} + 1)$, decrease as n_{diff} increases from 0 to 3); a complementary property is that n_{rel} modulates the effect of n_{diff} on $\bar{n}_{\text{tests}}$ (separations among the points in a column grow with n_{rel}).[8]

It is now instructive to jump ahead briefly to figure 9.7B, in which the data from figure 9.4A are replotted as a function of n_{rel}. For the R_{diff} responses in Bamber's experiment (1969), the structure of these RT data is remarkably similar to the behavior of the number of matching tests, according to the model we have been considering, from which we derived equation 9.2. In the next section we consider what this similarity might mean about the underlying process.

Anticipating section 9.3, note that for R_{same}, all relevant features are tested and all of these are matching tests, so that for any search path,

$$\bar{n}_{\text{tests}}(\text{same}) = \bar{n}_{\text{mtests}}(\text{same}) = n_{\text{rel}}. \tag{9.3}$$

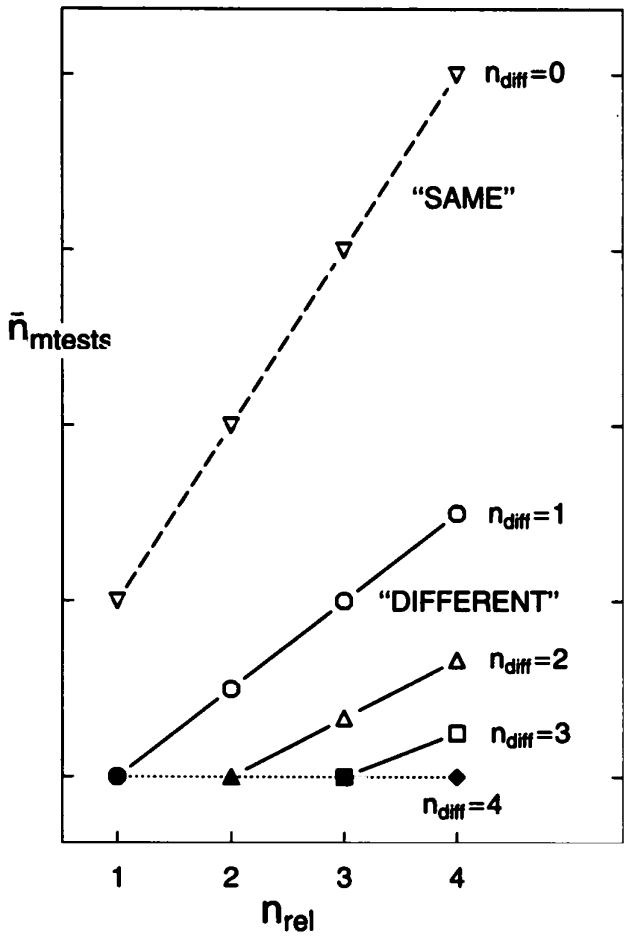

Figure 9.5
Mean number of matching tests, $\bar{n}_{mtests}$, expected from model with sequential self-terminating tests, as function of number of relevant attributes (or number of elements). Points connected by solid lines represent "different" trials, with $n_{diff} = 1, 2, 3, 4$, as described by combining equation 9.2 with equation 9.1. Filled points (and the dotted line) show cases where $n_{diff} = n_{rel}$. Points connected by the dashed line represent "same" trials, with $n_{diff} = 0$, as described by equation 9.3.

Thus $\bar{n}_{tests}$(same) increases linearly with n_{rel}, just as $\bar{n}_{tests}$(diff) does. Recall that even with fixed n_{rel} and n_{diff}, as illustrated in tables 9.1 and 9.2, a distribution of values of n_{tests}(diff) enters into the mean, $\bar{n}_{tests}$(diff). In contrast, n_{tests}(same) is unaffected by the search path; thus the "distribution" of n_{tests} that enters into the mean, $\bar{n}_{tests}$ contains just one value. The points connected by broken lines in figure 9.5 represent equation 9.3. From equations 9.2 and 9.3 you can see that the slopes of the four linear functions in the figure, for $n_{diff} = 0, 1, 2, 3$, (in units of the number of tests added for each unit increment in n_{rel}) are 1, 1/2, 1/3, and 1/4, respectively.

Comment 4: Testing as search. I mentioned above that we can think of the sequence of feature tests as a self-terminating search through

a set (of pairs of corresponding features or pairs of corresponding elements) that contains either one or more "targets" (mismatching pairs, $n_{diff} \geq 1$), or no targets (all pairs match, $n_{diff} = 0$). If the set contains n_{rel} members, and n_{diff} targets, and the targets are located randomly in the search path, then equation 9.2 applies to the "positive" (target-present, R_{diff}) trials. Similarly, equation 9.3 applies to the "negative" (target-absent, R_{same}) trials. Suppose there is exactly one target on positive trials ($n_{diff} = 1$), as in many memory-search and visual-search tasks (such as searching for a book in your personal collection), and the set of elements to be searched (books on the shelf) is of size s. We can set $n_{rel} = s$ in equations 9.2 and 9.3, and $n_{diff} = 1$ in equation 9.4, and have $\bar{n}_{tests}(positive) = (s + 1)/2$, while $n_{tests}(negative) = s$. For both positive and negative responses, the number of tests rises linearly with s, but when no target is present (a negative trial) the number of tests rises twice as rapidly with s. (Consider the effect of increasing s by two elements: $n_{tests}(negative)$ increases by 2, but $\bar{n}_{tests}(positive)$ by only 1.) This corresponds to the 2:1 slope ratio of the top two functions in figure 9.5, for $n_{diff} = 0$ and $n_{diff} = 1$.[9] (See Dosher, chap. 10, this volume, for discussion of the application of this result to visual search performance.)

9.2.3 Sequential Tests: Relation between the Number of Tests and Mean Reaction-Time

In the development thus far, we have seen predictions that the sequential model makes about n_{tests}. But because we measure reaction time, RT, and not the number of tests, we have to consider what the predictions for n_{tests} imply about RT. In doing so, we encounter three assumptions that are fundamental in many of the inferences that are made from patterns in RT data to the structure of underlying processes. Coloring my discussion of these assumptions is the remarkable similarity noted above for R_{diff} between the mean number of tests (figure 9.5) and $\overline{RT}$ (figure 9.7B). This similarity means that $\overline{RT}$ increases by a constant amount for each unit increase in $\bar{n}_{tests}$. In other words, the similarity suggests a linear relationship between $\overline{RT}$ and the number of tests.

Comment 5: Discovering assumptions. There is much to be gained by attempting to identify the assumptions on which our reasoning depends. Because commonly held assumptions are often only implicit in a theory, we may not appreciate their contributions to our thinking (consider, for example, the assumption of a "ballistic response process," discussed in section 9.2.4.6). In this chapter, I attempt to identify some of the assumptions that underlie much of the reasoning about multiattribute and multielement stimulus comparison; there

may be other interesting ones worth considering, of which I am unaware. There is much to be gained by making assumptions explicit. We might recognize an assumption as implausible. A required assumption might be true if we ran the experiment or analyzed the data in one way, but not another. Once an assumption is revealed, we may discover there is relevant evidence in some other empirical domain, or we might think of an experimental test of that assumption in isolation from others. While making assumptions explicit, we should consider the point made by Doucet and Sloep (1992, 285) that "no list of assumptions could ever be complete ... there may always be relevant factors that we simply have not thought of ... such is the fate of model-building and indeed of all science: there are no guarantees." They also point out, however, that violations of many of our assumptions are sufficiently improbable that there is little advantage in making them explicit; their example of such an assumption is "no tampering by intelligent mischievous aliens."

9.2.3.1 *The Contribution of Residual Operations to Reaction Time*

Between the presentation of the two patterns and the occurrence of the response is a sequence of hypothetical operations, shown in figure 9.6. One operation consists of the feature tests discussed above; **FT** refers to this testing process, and T_{ft} denotes its duration. There are also *residual operations* that probably include at least three distinct processes. Suppose RT is measured from presentation of the second pattern, as in Bamber's experiment, with plenty of time provided for encoding the first. One of

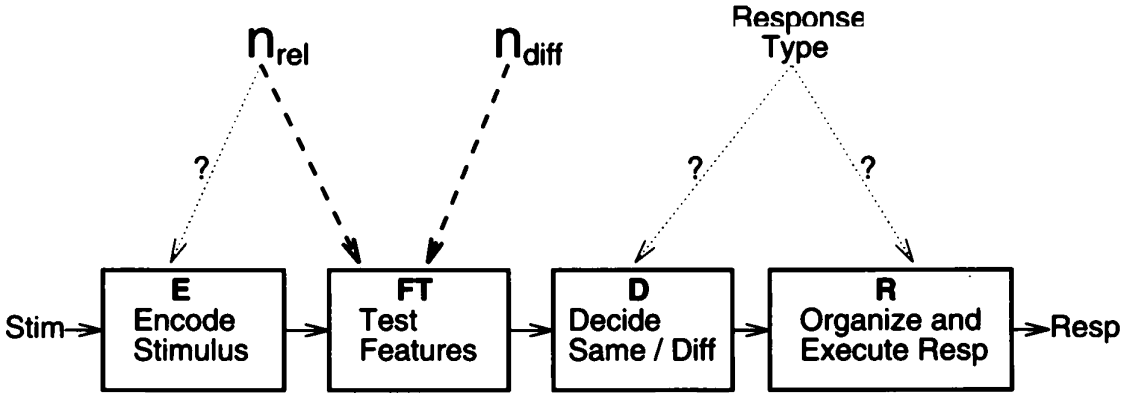

Figure 9.6
Flowchart of hypothetical processing stages between test stimulus ("stim") and "same" or "different" response ("resp") in Bamber's experiment (1969). Stage **FT** is the operation of central interest. The descending arrows represent possible influences on stage durations of factors $n_{rel} = 1, 2, 3, 4$; $n_{diff} = 0, 1, 2, 3, 4$; response type "same" ($n_{diff} = 0$) or "different" ($n_{diff} > 0$).

the residual operations is then **E**, with duration T_e, the encoding of the second pattern, the process that forms the representation of that pattern that is used by **FT**.[10] A second residual operation is **D**, with duration T_d, in which the decision is made as to whether the two patterns are the same or different, based on information furnished by **FT**. And a third residual operation is **R**, with duration T_r, which organizes and executes the response, based on the decision. I call these operations "residual," not because they are uninteresting, but because they are of secondary concern in the present context. The flowchart of figure 9.6 shows the three hypothetical residual operations, together with the feature testing operation, which together occupy the time between stimulus and response:

$$RT = T_e + T_{ft} + T_d + T_r. \tag{9.4}$$

Let α (alpha),[11] denote the mean total duration of the residual operations, $\alpha = \bar{T}_e + \bar{T}_d + \bar{T}_r$.

Stages. Because the feature-testing, decision, and response processes are *data-dependent*, with each making use of information provided by its predecessor, it is plausible that they are arranged in stages, with one process beginning when the preceding process ends. This is shown in the flowchart, in which time proceeds from left to right. From this *stages assumption*, together with the fact that the mean of a sum is equal to the sum of the means of the summands, $\text{mean}(X + Y) = \text{mean}(X) + \text{mean}(Y)$, it follows that $\overline{RT}$ is given by the sum, α, of the mean durations of the residual operations **E**, **D**, and **R**, plus the mean duration of the testing operation, $\bar{T}_{ft}$:

$$\overline{RT} = \alpha + \bar{T}_{ft}. \tag{9.5}$$

Selective influence. According to a *selective influence assumption*, the variables or *factors* under discussion (n_{rel} and $n_{diff} \geq 1$) influence only T_{ft}, and not α; that is, changes in the levels of these factors change the duration of only one of the four stages. This is shown in figure 9.6 by the heavy descending arrows from "n_{rel}" and "n_{diff}" to "**FT**." It is not obvious that such selective influence should obtain; for example, variation of n_{rel} could be imagined to influence an encoding operation, with encoding being more elaborate and taking longer when n_{rel} is greater; this could be especially important for Bamber's letter-string experiment, where the complexity of the display increases systematically with n_{rel}, and where an encoding process might operate on all n_{rel} of the letters before **FT** starts. The light descending arrow from "n_{rel}" to "**E**" in figure 9.6 reminds us to be alert for this possibility. Despite its plausibility, the possibility of a separate encoding process whose duration increases with n_{rel} is excluded from the sequential mechanism as it is developed below (following Bamber) to

explain RT data. Indeed, one of the great appeals of Bamber's model is the simplifying property that there is no n_{rel}-dependent process other than the one also influenced by n_{diff}. (In this regard the competing parallel-test model to be discussed below is less appealing.) Thus the effects of n_{rel} and $n_{diff} \geq 1$ on RT are mediated entirely by their effects on $\bar{n}_{tests}(diff)$, as expressed by equation 9.2. We shall see below that a test of this assumption provides strong support.

> *Comment 6: Decomposing a mechanism.* Dividing a mechanism into a part (such as **FT**) that is under study and that we attempt to manipulate, and the remainder (which we hope is not influenced by the manipulations) is, of course, a common research strategy. In the present case the decomposition depends on assumptions (stages and selective influence) that have not been explicitly tested in the object-comparison experiment.[12]

A second possible complication arises when we attempt to extend our account to cover R_{same} as well as R_{diff}, by including $n_{diff} = 0$ in the range of n_{diff}. Because the decision and response depend on whether a mismatch has been discovered, the outcome of **FT** may influence **D** and **R**, and hence influence α. This possibility is expressed in figure 9.6 by the inclusion of "response type" (R_{same} or R_{diff}) as a factor that may influence the durations of **D** and/or **R**. Also to express this possibility, I distinguish between the αs in the two cases, and use α_{same} for R_{same} and α_{diff} for R_{diff}:

$$\overline{RT}_{diff} = \alpha_{diff} + \bar{T}_{ft}(diff), \qquad (n_{diff} > 0); \tag{9.6}$$

$$\overline{RT}_{same} = \alpha_{same} + \bar{T}_{ft}(same); \qquad (n_{diff} = 0). \tag{9.7}$$

Hence any changes induced in $\overline{RT}_{diff}$ by variations in n_{rel} and n_{diff} are direct reflections of and are equal to the effects of those variations on $\bar{T}_{ft}$, as are changes in $\overline{RT}_{same}$ induced by n_{rel}. Our problem of relating n_{tests} (or n_{rel} and n_{diff}) to $\overline{RT}$ is therefore reduced to the problem of relating them to $\bar{T}_{ft}$.

> *Comment 7: Unitary factors.* This potential change in the pattern of influence of n_{diff} as its range is extended from $n_{diff} \geq 1$ to include $n_{diff} = 0$ illustrates the importance of considering whether a factor might or might not be *unitary*, in using its effects on RT to analyze mental processes. We shall see that $n_{diff} \geq 1$ is probably unitary, in the sense that each increment, from 1 to 2, from 2 to 3, and from 3 to 4, appears to influence the same process and leave the same other processes invariant. In contrast, we shall find that the increment from 0 to 1 departs from this uniform pattern, so $n_{diff} \geq 0$ is not unitary. (Can you see why?)

In the class of models we are considering, testing is assumed to be sequential. This means that the tests are carried out one after another, such that one test begins when the preceding one ends. (This is another instance of a stages assumption, here applied to the internal structure of a process already defined as a stage; we might describe the tests as substages of the feature-testing stage.) It follows that T_{ft} is a sum of the durations of the n_{tests} feature tests. To make use of this summation property, we need an additional assumption.

Pure insertion. According to a *pure-insertion assumption,* the duration of a particular test is independent of its context—that is, of the particular other tests in which it is embedded, or the number of such tests. The "insertion" is "pure" in the sense of "only": the particular test is inserted into the sequence of other tests and other operations without changing any of them. (As an analogy, suppose you are about to produce four short documents, one after another, on a printer that is well supplied with paper, and you insert a fifth document into the queue.)

9.2.3.2 Implications of Four Constraints on Test Durations

The relation between T_{ft} and n_{tests} is straightforward if four simplifying constraints apply to the durations of feature tests and residual operations. Let us therefore begin by examining the resulting model and how well it accounts for Bamber's data (1969). In section 9.2.4 we consider whether the predictions of the model change as these constraints are relaxed, and if so, how. It is important to consider this because all four of the constraints are implausible, except, perhaps, as approximations. (I use "constraint" to denote assumptions that are less fundamental than what I have called "defining properties." They might also be called "side conditions." We would not discard a theory because of the failure of a constraint; the alteration of a theory by relaxing a constraint would not be an essential alteration.)

> *Constraint 1:* A particular test has the same duration from one occasion (trial) to the next;
> *Constraint 2:* The residual operations for "same" and "different" responses have the same durations: $\alpha_{same} = \alpha_{diff} = \alpha$;
> *Constraint 3:* Tests that lead to matches and to mismatches of an attribute have the same mean duration;
> *Constraint 4:* Tests of different attributes have the same mean duration.[13]

Given these four constraints, the duration of a test is a fixed constant, θ (theta). It follows that $\bar{T}_{ft} = \theta \bar{n}_{tests}$, and therefore, from equation 9.2, that

$$\overline{RT}_{\text{diff}}(n_{\text{rel}}, n_{\text{diff}}) = \alpha + \theta\bar{n}_{\text{tests}}(\text{diff})$$
$$= \alpha + \theta\left(\frac{n_{\text{rel}} + 1}{n_{\text{diff}} + 1}\right), \qquad (1 \leq n_{\text{diff}} \leq n_{\text{rel}}); \tag{9.8}$$

and it follows from equation 9.3 that

$$\overline{RT}_{\text{same}}(n_{\text{rel}}) = \alpha + \theta\bar{n}_{\text{tests}}(\text{same}) = \alpha + \theta n_{\text{rel}}, \qquad (n_{\text{diff}} = 0). \tag{9.9}$$

According to the model, the same linear function, $\alpha + \theta\bar{n}_{\text{tests}}$, should describe the increase in mean reaction-time with $\bar{n}_{\text{tests}}$ for both R_{same} and R_{diff}. (For both responses, any variation in n_{tests} is due to variation in the number of matching tests.) Furthermore, because the effects of n_{rel} and n_{diff} on $\overline{RT}_{\text{diff}}$ are mediated by their effects on $\bar{n}_{\text{tests}}(\text{diff})$ (as described in equation 9.2), and because $\overline{RT}_{\text{diff}}$ is a linear function of $\bar{n}_{\text{tests}}(\text{diff})$, properties 1–4 of $\bar{n}_{\text{tests}}(\text{diff})$ described in section 9.2.2.3 carry over to $\overline{RT}_{\text{diff}}$. One consequence is that rather than being invariant over levels of n_{rel}, the effect of n_{diff} on $\overline{RT}_{\text{diff}}$ is modulated by n_{rel}, and vice versa. (To demonstrate this, use equation 9.8 to compare the change in $\overline{RT}_{\text{diff}}$ produced by increasing n_{diff} from 1 to 2—a measure of the effect of n_{diff}—when $n_{\text{rel}} = 2$ and when $n_{\text{rel}} = 4$.) That is, the factors n_{diff} and n_{rel} *interact*, rather than having additive effects on $\overline{RT}_{\text{diff}}$. (We shall see in chap. 14, this volume, that interactions play an important role in concluding that two such factors influence the same stage of processing, here, **FT**, in situations where we are considering a less explicit model than Bamber's.)

9.2.4 Sequential Tests: Application to Letter-String Data

9.2.4.1 The Fully Constrained Model

Bamber (1969) fitted the sequential model with constraints 1–4 to the $\overline{RT}_{\text{diff}}$ data from his experiment. The data here (see figure 9.4A) consist of ten mean *RT*s. "Fitting" the model to the data consists of finding those values of the two "free parameters" α and θ that minimize some measure of the discrepancies (deviations) between the ten data points and the values given by equation 9.8. In this case the measure used (method of "least squares") was the sum of the squared deviations, and the estimates were $\hat{\alpha} = 323.8$ ms and $\hat{\theta} = 60.5$ ms/test. (A parameter with a "hat" denotes an estimate of that parameter obtained from data. See Dosher, chap. 10, this volume, for a discussion of parameter estimation and the fitting of models to data.) The model fitting was allowed to depend on only the $\overline{RT}_{\text{diff}}$ data because preliminary comparison of data and model suggested that the sequential model would not be able to account for the behavior of both $\overline{RT}_{\text{diff}}$ and $\overline{RT}_{\text{same}}$, and because the greater complexity of $\overline{RT}_{\text{diff}}$ is more challenging for any model.

Three ways of viewing the relation between the fitted model (dotted lines) and the data are shown in figure 9.7. In panel A the data are shown in the same way as in figure 9.4A, but here the corresponding fitted values for the model have been added. The model fits the $\overline{RT}_{\text{diff}}$ data well but fails miserably in fitting the $\overline{RT}_{\text{same}}$ data. One possible difficulty revealed by this figure is the tendency for the effect of n_{diff} to be too large, relative to the model, as it increases from 1 to 2, and too small as it increases from 2 to 3, and from 3 to 4. That is, the data show greater diminishing returns than the model does. More data are probably needed to decide whether this is a true discrepancy.

In panel B, the presentation is reorganized so as to make it easier to see any systematic violations of three quantitative properties of the model (properties that follow from three of those discussed in relation to figure 9.5):

1. $\overline{RT}_{\text{diff}}$ for each n_{diff} should increase linearly with n_{rel}, as shown by the rising dotted lines;
2. The slopes of these linear functions should be proportional to $1/(n_{\text{diff}} + 1)$, as shown by the relations among those lines;
3. If $n_{\text{diff}} = n_{\text{rel}}$, $\overline{RT}_{\text{diff}}$ should not be affected by n_{rel}, as shown by the horizontal dotted line.

Whereas the deviations from the first two of these properties seem unsystematic, deviations from the third (expressed by the relation between the filled points and the horizontal dotted line) hint at a tendency for $\overline{RT}$ to increase with n_{rel}, that is, with the number of displayed letters. However, this trend is not consistent, even for the filled data points. Furthermore, the deviations of the other (unfilled) data points from the fitted values are not consistent with any tendency for $\overline{RT}$ to increase with n_{rel} more than expected from the model. Such a discrepancy would be expected if n_{rel} influenced the duration of a separate encoding process (E in figure 9.6), as discussed in section 9.2.3.1. Although any such increase appears to be relatively small in this experiment, larger effects of this kind have been seen in other experiments (see Bamber 1972; Eichelman 1970).

Comment 8: Measuring systematic deviation. A more quantitative way to determine whether n_{rel} influences the duration of a separate encoding process is to generalize equation 9.8 by adding a term to express such an effect. Let us approximate such a hypothetical effect on the duration of the encoding stage by an increase of ε (epsilon) ms for each displayed element. Equation 9.8 then becomes

$$\overline{RT}_{\text{diff}} = \alpha + \varepsilon n_{\text{rel}} + \theta\left(\frac{n_{\text{rel}} + 1}{n_{\text{diff}} + 1}\right), \qquad (1 \leq n_{\text{diff}} \leq n_{\text{rel}}). \tag{9.10}$$

When this more general model is fitted by least squares, the estimate of the added parameter is a negligible $\hat{\varepsilon} = 1.1$ ms per element, further evidence that the relation between the effects of n_{rel} and n_{diff} on $\overline{RT}_{diff}$ is accurately described by Bamber's model of the **FT** stage.

Another discrepancy that is especially noticeable in panel B is that the separation between the data points for $n_{diff} = 1$ and $n_{diff} = 2$ is too large, relative to the model, while the effects of n_{diff} from 2 to 3 to 4 are too small. Indeed, this alternative description of the deviations of the $n_{diff} = n_{rel}$ data is worth keeping in mind when seeking an explanation for them.

In panel C, the relation between model and data is shown in a third way, so as to make it easier to see any systematic discrepancies from the linear relation (equations 9.8 and 9.9) expected between $\bar{n}_{tests}$ (as given by equations 9.2 and 9.3) and $\overline{RT}$. The R_{diff} data not connected by any lines cluster close to the linear function (dotted line) that was fitted to them; there is no hint of any curvilinear tendency, as we might expect if mean test duration increased or decreased with the total number of tests. However, the R_{same} data, shown by the four triangles connected by line segments, deviate dramatically from the dotted line; it is clear that no single linear function could possibly provide a good description of both the R_{diff} and the R_{same} data. The most obvious difficulty for the R_{same} data is that the RTs expected from the model are much too long. For example, the rightmost $\overline{RT}_{same}$ value is 391 ms; the value expected by the model is 564 ms, or 173 ms greater.

To what extent do the four constraints adopted above contribute to either the success or failure of the attempt to explain the letter-string data? (If a theory fits a set of data only under a set of implausible side conditions, then this should probably be taken as evidence *against* the theory!) Let us consider what happens as we relax the constraints. We shall see that as we relax the first three constraints, the properties of the model we have considered remain about the same, but that as we relax the fourth constraint, some difficulties are revealed.

9.2.4.2 Relaxing Constraint 1: Allowing Variable Test Durations

According to constraint 1, the duration of the test of a particular feature is fixed from one occasion to the next, rather than varying. Among the properties of sequential models that render them especially easy to work with, one is that a change from fixed, deterministic process durations to the more realistic variable process durations has no effect whatsoever on mean reaction time $\overline{RT}$.[14] In contrast, as we shall see in sections 9.2.5 and 9.3, for parallel processes the variability of process durations is relevant, even if we limit our interest to properties of the RT mean. This is unfortunate

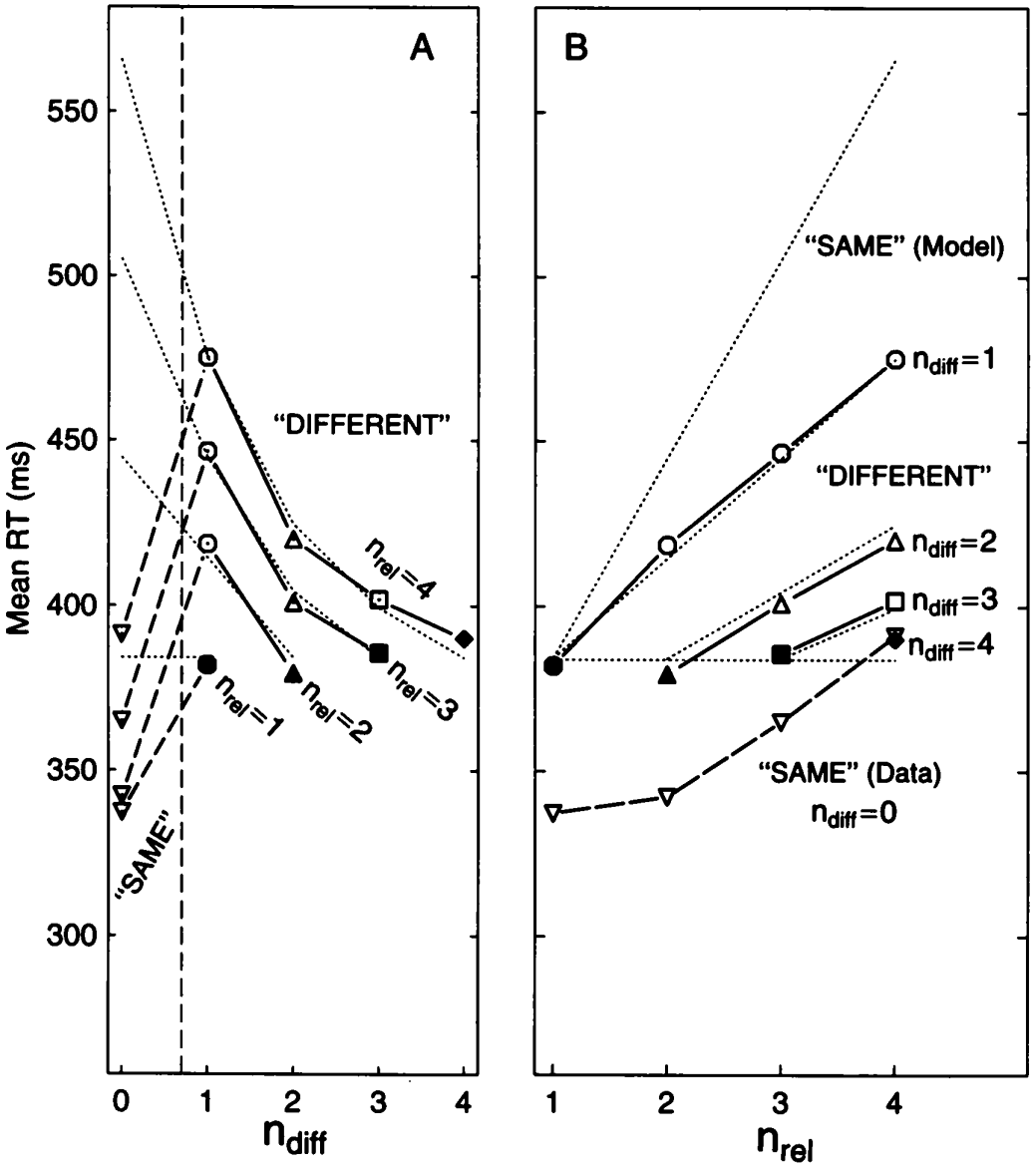

Figure 9.7

Three views of $\overline{RT}$s from Bamber's experiment (1969) versus fitted values for variant of sequential-test model expressed by equations 9.2, 9.3, 9.8, and 9.9. Parameter estimates $\hat{\alpha} = 323.8$ ms and $\hat{\theta} = 60.5$ ms per test were chosen so as to minimize the sum of squared deviations of the $\overline{RT}_{diff}$ values from the corresponding model values given by equation 9.8; $\overline{RT}_{same}$ values played no role in the parameter estimation. The fourteen data values are shown by circles, triangles, squares, and diamonds, at the same heights in the three panels; model values are connected by dotted lines. Data points are filled for the four cases where $n_{diff} = n_{rel}$ (= 1, 2, 3, 4). In panel A, $\overline{RT}$ is plotted for each value of n_{rel} (number of displayed letters) as a function of n_{diff} (number of letters in corresponding positions that differ), as in figure 9.4A. In panel B, $\overline{RT}$ for each value of n_{diff} is plotted as a function of n_{rel}. Rounded slopes of the fitted functions for $n_{diff} = 0$, 1, 2, and 3 are $\hat{\theta} = 61$, $\hat{\theta}/2 = 30$, $\hat{\theta}/3 = 20$, and $\hat{\theta}/4 = 15$ ms per displayed letter, respectively. Data values for R_{diff} fall very close to their model values, but data values for R_{same} (downward-pointing triangles) deviate markedly in

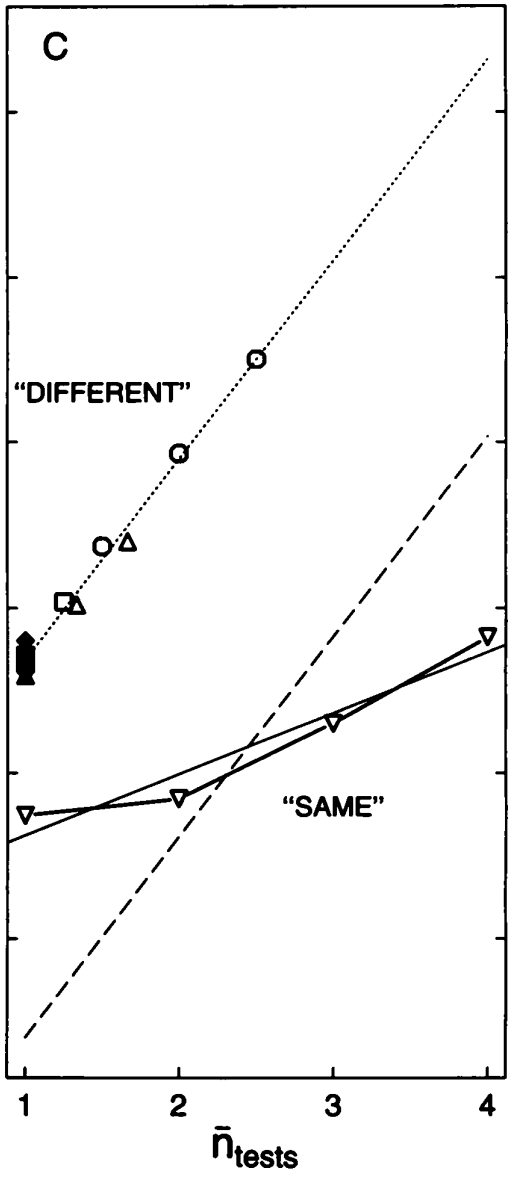

Figure 9.7 (cont.)
both slope and absolute value from their model values (uppermost dotted line). In panel C, $\overline{RT}$ is plotted as a function of the predicted $\bar{n}_{tests}$ described by equations 9.2 and 9.3. The upper dotted line is $\overline{RT} = 323.8 + 60.5\bar{n}_{tests}$; given constraints 2 and 3, both $\overline{RT}_{diff}$ and $\overline{RT}_{same}$ should fall on this line. If we relax constraints 2, 3, or both, this permits $\overline{RT}_{same}$ to fall on a different line with the same slope. The parallel dashed line has a smaller intercept but the same slope: $\overline{RT} = 207.7 + 60.5\bar{n}_{tests}$. The unbroken line is the best-fitting line for R_{same}, $\overline{RT}_{same} = 312.9 + 18.4\bar{n}_{tests}$. The plotted data (with fitted values in parentheses) for increasing levels of n_{rel} are as follows: $n_{diff} = 0$: 337.3 (384.3), 342.4 (444.8), 365.1 (505.3), 391.2 (565.8); $n_{diff} = 1$: 382.0 (384.3), 418.6 (414.6), 446.6 (444.8), 475.1 (475.1); $n_{diff} = 2$: 379.6 (384.3), 401.1 (404.5), 420.1 (424.6); $n_{diff} = 3$: 385.7 (384.3), 401.9 (399.4); $n_{diff} = 4$: 390.1 (384.3).

because, as a result, predictions about the mean durations of parallel mechanisms are much less general.

9.2.4.3 *Relaxing Constraint 2: Allowing Unequal Residual Durations for "Same" and "Different" Responses*

According to constraint 2, the residual operations for R_{same} and R_{diff} have the same durations. In contrast, the processes of decision and response organization, which presumably follow the sequential testing process and make use of the information it provides, would plausibly have different durations for R_{same} and R_{diff}. One argument favoring such duration differences, for example, is based on evidence from related experiments suggesting that we can manipulate the bias for "same" versus "different" responses, $\overline{RT}_{diff} - \overline{RT}_{same}$, without influencing $\overline{T}_{ft}$, by varying the relative occurrence frequency of the two responses.[15] Given that the residual durations can differ, there is no reason to assume that they are equal in any particular set of conditions. Permitting α_{same} and α_{diff} to be assigned unequal values leads to a version of the sequential model expressed by equations 9.11 and 9.12, more general than equations 9.8 and 9.9:

$$\overline{RT}_{diff} = \alpha_{diff} + \theta\bar{n}_{tests}(diff), \qquad (n_{diff} > 0); \tag{9.11}$$

$$\overline{RT}_{same} = \alpha_{same} + \theta n_{tests}(same), \qquad (n_{diff} = 0). \tag{9.12}$$

This generalization does not influence how well the model can fit the R_{diff} data alone, but it does slightly improve the model's ability to account jointly for the R_{same} and R_{diff} data. Recall that $\hat{\alpha}_{diff} = 324$ ms. If we let $\alpha_{same} = \alpha_{diff} - 114$ ms $= 210$ ms, then the fitted values for $\overline{RT}_{same}$ in figure 9.7C fall on the lower dashed line rather than the upper dotted one.

9.2.4.4 *Relaxing Constraint 3: Allowing Unequal Durations of Matches and Mismatches*

This constraint is relevant because we want to fit the model to $\overline{RT}_{same}$ as well as $\overline{RT}_{diff}$. Suppose β (beta) and γ (gamma) are the durations of matching and mismatching tests, respectively. According to constraint 3, they are equal: $\gamma - \beta = 0$. Suppose, instead, they are permitted to differ, and that the difference between them is $\gamma - \beta = \delta$ (delta), where δ may be positive or negative. Because $\overline{RT}_{diff}$ contains a mismatch, whereas $\overline{RT}_{same}$ does not, a positive value of δ could contribute to the 114 ms $\alpha_{diff} - \alpha_{same}$ difference noted above. According to the model, RT_{diff} is based on a sequence that contains one mismatching test and from zero to $n_{rel} - 1$, (or an average of $\bar{n}_{tests}(diff) - 1$) matching tests. It follows that

$$\begin{aligned}\overline{RT}_{diff} &= \alpha_{diff} + \gamma + \beta[\bar{n}_{tests}(diff) - 1] \\ &= (\alpha_{diff} + \delta) + \beta\bar{n}_{tests}(diff) = \alpha'_{diff} + \beta\bar{n}_{tests}(diff).\end{aligned} \tag{9.13}$$

By replacing α_{diff} by $\alpha'_{diff} = \alpha_{diff} + \delta$, we absorb the constant δ into the first term. Because α_{diff} incorporates the durations of other unknown processes, replacing it by α'_{diff} does not add indeterminacy. We can thus continue to use equation 9.11, replacing θ by β to remind ourselves that what varies from trial to trial and condition to condition is the number of feature tests that lead to a match. Also according to the model, RT_{same} is based on a sequence that contains $n_{tests}(same)$ matching tests, so 9.12 becomes

$$\overline{RT}_{same} = \alpha_{same} + \beta n_{tests}(same). \tag{9.14}$$

9.2.4.5 Relaxing Constraint 4: Allowing Unequal Mean Test-Durations for Different Attributes

We have seen that the predictions of the model described by equation 9.8 for $\overline{RT}_{diff}$ are not altered when we relax the first three constraints. Constraint 4 is more critical, however. Imagine deciding whether the following pairs of the patterns in figure 9.1 are the same: when size is relevant: (a) and (e); when direction is relevant: (a) and (c); when shape is relevant: (a) and (b). In each case, only the values of the relevant attribute differ between the elements of the pair; presumably only the relevant attribute is tested. According to constraint 4, the tests of different features (or letter locations) have the same durations. Does this seem plausible for these patterns?

> *Comment 9: Forcing constraint 4.* Before considering the effects of relaxing this constraint, it is worth noting ways in which we might be able to cause it to be at least approximately satisfied. In one approach it should be possible to adjust either the choice of attributes or the difference between the values of attributes (their discriminability), so as to satisfy the constraint approximately. With only one attribute at a time defined as relevant, we could adjust the difference between its values so as to equate $\overline{RT}_{same}$ across attributes; for geometric patterns, it would seem unlikely that the constraint would be satisfied without such efforts. (At the very least, this approach permits us to test how well constraint 4 is approximated.)[16] Another approach would be to arrange that each pattern consists of a set of n_{rel} "pieces," each providing a value of the *same* attribute. For example, each piece might be a shape—either a square or a circle—and the shapes might be laid out in a row to make up the pattern. R_{same} would be correct if the shapes in corresponding positions in the pair of patterns were identical. Bamber's letter-string experiment (1969) can be regarded as an example of this approach. Consider the test of whether a letter in a position matches or mismatches a previously seen letter in the corresponding position. Insofar as the test

duration is approximately independent of position (which is surely not always the case and would have to be carefully tested), then the constraint would be approximately satisfied; in an experiment like Bamber's, one way to help achieve this is to arrange, as he did, that the possible letters have the same chance of appearing in each position.

Suppose there are systematic differences among test durations for the different attributes in a pattern experiment (or the different positions in a letter-string experiment), thus violating constraint 4. In general, this will make the relation between $\bar{n}_{\text{tests}}$ and $\overline{RT}$ indeterminate, because it now matters *which* features (or letter positions) are tested, and not simply how many. Thus, in general, the quantitative relationship expressed in equation 9.11 will not hold exactly. This leads to three questions:

1. Are there any conditions under which equation 9.11 will still be valid?
2. Is there some statement weaker than equation 9.11 that we can make about the effects of n_{diff} and n_{rel} on $\overline{RT}$?
3. How large will the deviation from equation 9.11 be for differences of plausible magnitude among test durations; that is, how sensitive is equation 9.11 to such violations?

In this chapter I comment on questions 1 and 2 (to which the answers are "yes"), but, because of the complexity of the considerations, not on question 3.

The features actually tested on a trial depend on two variables: which features match and mismatch (controlled by the experimenter, and determining the *trial type*, or the column in table 9.3) and the order in which

Table 9.3
The sequence of feature tests as a function of mismatching feature(s) (column) and search path (row) when $n_{\text{rel}} = 3$ and $1 \leq n_{\text{diff}} \leq 3$

	$sn_{\text{diff}} = 1$			$n_{\text{diff}} = 2$			$n_{\text{diff}} = 3$
Mismatching feature(s)	A	S	D	A,S	A,D	S,D	A,S,D
Search path							
A→S→D	**A**	*a*,**S**	*a*,*s*,**D**	**A**	**A**	*a*,**S**	**A**
A→D→S	**A**	*a*,*d*,**S**	*a*,**D**	**A**	**A**	*a*,**D**	**A**
S→A→D	*s*,**A**	**S**	*s*,*a*,**D**	**S**	*s*,**A**	**S**	**S**
S→D→A	*s*,*d*,**A**	**S**	*s*,**D**	**S**	*s*,**D**	**S**	**S**
D→S→A	*d*,*s*,**A**	*d*,**S**	**D**	*d*,**S**	**D**	**D**	**D**
D→A→S	*d*,**A**	*d*,*a*,**S**	**D**	*d*,**A**	**D**	**D**	**D**

features are tested (the search path, controlled by the subject, and determining the row in table 9.3). Information in the first row of table 9.3 is also contained in the first three columns of table 9.1.

Given the column (associated with a particular trial type), the six test sequences correspond to the six possible search paths, and are given by the six entries in that column. The test sequences are all composed of one or more of the six tests: *a*, *s*, *d*, **A**, **S**, and **D**. Suppose each of these tests has a different mean duration associated with it: β_A, β_S, and β_D, for tests of features that match (nontargets), and γ_A, γ_S, and γ_D, for tests of features that mismatch (targets). Then, for example, if A and D are the mismatching features (sixth column), and the search path is S → D → A (fourth row), then the test sequence is *s*, **D**, and $T_{ft} = \beta_S + \gamma_D$.

If the six test durations are permitted to differ freely and we know nothing about the search path, it should be evident that we can say very little about the relation between $\bar{T}_{ft}$ and n_{diff}. In contrast, if we make the strong assumption that the search path is random (i.e., that the six paths are equally likely), we can say a great deal. Given this assumption, if the experimenter arranges that the three columns of table 9.3 for $n_{diff} = 1$ (and the three columns for $n_{diff} = 2$) occur with equal frequency—trial types are *balanced*—then the eighteen test sequences for $n_{diff} = 1$ (and the eighteen for $n_{diff} = 2$) occur equally often. Let the mean of β_A, β_S, and β_D be $\bar{\beta}$ and the mean of γ_A, γ_S, and γ_D be $\bar{\gamma}$. Averaging the durations of these sequences we get, for $n_{diff} = 1$, $\bar{T}_{ft} = \bar{\gamma} + \bar{\beta}$, and for $n_{diff} = 2$, $\bar{T}_{ft} = \bar{\gamma} + \frac{1}{3}\bar{\beta}$; for $n_{diff} = 3$ we average over the six possibilities to get $\bar{T}_{ft} = \bar{\gamma}$. These predictions are summarized in table 9.4.

> *Comment 10: Virtues of balance.* This illustrates the simplification that often occurs when we can assume or impose equal frequency on a set of events (that is, balance the set). An alternative to equalizing frequencies among the columns in each set of three is to use an ordinary mean of each set of three column means. Substitution of such "statistical balancing" for experimental balancing can be helpful in designing experiments and analyzing data.

Table 9.4
Mean number of tests and mean duration of the feature-testing process as a function of n_{diff} ($1 \leq n_{diff} \leq 3$) when $n_{rel} = 3$, trial types are balanced, and search paths are equiprobable

n_{diff}	$\bar{n}_{tests}$	$\bar{T}_{ft}$
1	2.00	$\bar{\gamma} + \bar{\beta}$
2	1.33	$\bar{\gamma} + \frac{1}{3}\bar{\beta}$
3	1.00	$\bar{\gamma}$

Thus, as n_{diff} is varied, the behavior of $\bar{T}_{ft}$ mirrors the behavior of $\bar{n}_{tests}$ (see also table 9.1; the time reduction from $n_{diff} = 1$ to $n_{diff} = 2$ is twice as great as the reduction from $n_{diff} = 2$ to $n_{diff} = 3$), and equation 9.11 is satisfied. One answer to question 1, then, is that equation 9.11 is still valid when the search path is random. What about the data? The average of the six data sets for pattern comparisons shown in figure 9.3A is roughly consistent with the 2:1 ratio: the decrease in $\overline{RT}$ from $n_{diff} = 1$ to $n_{diff} = 2$ is twice as great as the decrease from $n_{diff} = 2$ to $n_{diff} = 3$. And the same is true for the letter-string data for $n_{rel} = 3$, as shown in figure 9.7B.

But it seems to me implausible that the alternative possible search paths are equiprobable; a subject might use a fixed path (such as fastest to slowest test, or leftmost letter to rightmost letter), or might vary the path but with unequal frequencies over the set of possible paths.[17] Suppose the search path is fixed; in particular, suppose it is the first one given in table 9.3 (the conclusions below, however, hold for all search paths, and hence for any mixture of paths over trials). Let $\bar{T}_{ft}(n_{diff})$ be the average over the trial types (columns of the table) for a particular value of n_{diff}. Then, using the entries in the first row in each of the three sections of the table, we get

$$\bar{T}_{ft}(1) = \tfrac{1}{3}(\gamma_A + \beta_A + \gamma_S + \beta_A + \beta_S + \gamma_D),$$

$$\bar{T}_{ft}(2) = \tfrac{1}{3}(\gamma_A + \gamma_A + \beta_A + \gamma_S), \text{ and}$$

$$\bar{T}_{ft}(3) = \gamma_A = \tfrac{1}{3}(\gamma_A + \gamma_A + \gamma_A).$$

It follows that

$$\bar{T}_{ft}(1) - \bar{T}_{ft}(2) = \tfrac{1}{3}(\beta_A + \beta_S + \gamma_D - \gamma_A), \tag{9.15}$$

and

$$\bar{T}_{ft}(2) - \bar{T}_{ft}(3) = \tfrac{1}{3}(\beta_A + \gamma_S - \gamma_A). \tag{9.16}$$

Now we can address question 2. With no constraints on the test durations, we cannot predict even the signs of the $\bar{T}_{ft}$ differences given by equations 9.15 and 9.16. The theory is too weak, in the sense that it is consistent with too large a range of data patterns.[18] The sequential-test theory predicts very little when we relax constraint 4 and are unwilling to assume equiprobable search paths.

> *Comment 11: Paradox of testability.* This conclusion, which surprised me, illustrates the possibility that without further elaboration of an interesting theory (in the form of added assumptions or constraints) to specify it more precisely, the theory may not be easily testable. In this sense, the theory is too weak. Yet if the elaborated, stronger

theory (sometimes called a "model") fails, it may not be easy to know whether it is the theory or the elaboration that is at fault. Happily, if the stronger, elaborated theory succeeds, then the weaker, more general theory gains support, a fortiori.

Are there alternative plausible constraints that would make the theory testable? One possibility is to assume that the range of test durations is limited, such that the longest test duration is less than twice as great as the shortest. In this case, differences 9.15 and 9.16 are both positive, so the ordinal relations found in the data are also predicted by the model. That is, an increase in n_{diff} causes a reduction in $\overline{RT}_{diff}$. Another possibility with even stronger consequences is to assume that $\beta_A - \gamma_A = \beta_S - \gamma_S = \beta_D - \gamma_D = \lambda$, namely, that the *difference* between a matching and mismatching test of a feature is the same for all features, even though test durations can differ from one feature to another. It can then be shown, by replacing γ_A by $\beta_A - \lambda$ and so on in equations 9.15 and 9.16, that $\overline{T}_{ft}(1) - \overline{T}_{ft}(2) = \frac{1}{3}(\beta_S + \beta_D)$, and $\overline{T}_{ft}(2) - \overline{T}_{ft}(3) = \frac{1}{3}\beta_S$. That is, not only are both differences positive, but the first difference is greater than the second, as observed in the data (figures 9.3A and 9.4A). The effect of n_{diff} on $\overline{RT}$ is decelerating, though we cannot say by how much; we are left with a qualitative, ordinal prediction, not a quantitative one.

We have seen that relaxation of constraint 4 weakens the model sufficiently to add serious complications to its evaluation, complications that can be diminished either by making the dubious assumption of equiprobable search paths, or by invoking other constraints on test durations. The goal of creating conditions under which constraint 4 is approximately satisfied therefore becomes appealing. It is perhaps because the conditions of Bamber's letter-string experiment (1969) approximately satisfy the constraint—with letter-test duration approximately independent of letter position—that we see such excellent quantitative agreement between data and model.

9.2.4.6 Implications of a Nonballistic Response Process

In comment 5, I suggested that it was valuable to bring out implicit assumptions that guide our thinking. The assumption of a ballistic response process is an example. A process P_2 is *ballistic* if, once triggered by a process P_1, P_2 can no longer be controlled or influenced by P_1 (just as an unguided projectile, once launched, can no longer be controlled by the gunner). The way many of us think about the determinants of a reaction, and hence of the RT, is based on the idea of a single initiating event, such as the detection of a light flash. However, for stimuli with multiple attributes or components, where more than one target is present, the first target detection may not be the only one. After one event triggers the response

process, a second event may occur that would have been capable of the same triggering had it occurred alone. Although it is far easier to make quantitative predictions for a process if we can assume that only the first event has any influence, plausibility and some evidence (of "coactivation" of responses by multiple targets; see, for example, Miller 1982; Giray and Ulrich 1993) argue that the other alternative merits serious consideration. (This alternative is sometimes called "pipelining" to capture the idea of multiple signals passing through the same mechanism.) A self-terminating search is typically modeled as a process in which the tests end when a target (here, a mismatch) is found and the response is initiated. An alternative possibility, however, is that after a target is found, the testing process continues. Such a continuing testing process might influence RT.

Suppose two features differ, so that a second target is present. If the response process were ballistic, then the second mismatch would have no effect on RT_{diff}; if it were not, a second mismatch might shorten RT_{diff}. (Think of a sprinter, having started running in response to the starter pistol, being spurred to run faster by enthusiastic cheers from the spectators, or by the discharge from a second starter pistol.) An increase in n_{diff} would then have two effects. First, as described by the sequential model, it would reduce the mean number of tests before the first mismatch. But second, any subsequent matches that occur soon enough would also facilitate the response triggered by that first mismatch. Given this second effect of n_{diff}, how would the data deviate from the model?

The model prescribes relationships between $\bar{n}_{tests}$ and n_{rel} when $n_{diff} = 1$, and between $\bar{n}_{tests}$ and n_{diff} for fixed n_{rel} (see equation 9.2). The facilitation effect would not alter the former, but it would increase the effects of n_{diff}: in consequence, the data in figure 9.7A should fall increasingly below the predictions of the model as n_{diff} increases. The fact that this does not occur thus argues against the facilitation effect, and indicates that the response process is ballistic for the letter-string task, that testing does not continue beyond the first mismatch, or both. (This is an example of using a model as a baseline, mentioned in section 9.1.5. Fitting a model for the first effect of n_{diff} helps to test more sensitively for the presence of its second effect.)

9.2.4.7 Status of the Squential-Test Model

The mean RTs for R_{diff} from Bamber's letter-string experiment (1969) are beautifully described by the sequential-test model. The same simple process accounts for the effects of n_{rel}, of n_{diff}, and of the modulation of the effect of one by the other (their interaction); there is no need to postulate an effect of n_{rel} on any process other than the one also influenced by n_{diff}. An advocate of this model might have two concerns, however. First, there is a hint of a systematic deviation of data from model when $n_{diff} = n_{rel}$,

which should be investigated further when more data are available. Second, testability of the model depends heavily on constraint 4, which may not always be easy to satisfy.

Insofar as a model with sequential self-terminating tests is supported, this in turn indicates that analysis of visual forms into their component features plays an important role in their comparison, at least under conditions where the attributes (or elements) are well defined. If we accept the model, we can proceed to use it to estimate $\bar{\beta}$, the average time for testing a single feature pair (or letter pair) in the case of a match. From table 9.4 we see that the effect on $\overline{RT}$ of changing from $n_{\text{diff}} = 3$ to $n_{\text{diff}} = 1$ provides an estimate of this time. Figure 9.3A shows that this estimate ranges from about 50 ms to about 140 ms, depending on the experiment and the degree of practice. The corresponding value in figure 9.7C, where a "feature" corresponds to a single letter in a string of letters, is 60 ms.

9.2.5 Parallel Tests: Defining Properties

In this section we consider whether a parallel-testing process could underlie the R_{diff} data of figures 9.3 and 9.4. Given the success of the sequential-test theory for the data, you may wonder why we should ask whether an alternative theory can explain them. There are at least four reasons. First, because the brain appears capable of parallel processing, and subjects report being unaware of carrying out sequential tests, the idea that the testing process is sequential seems implausible. Second, it has been found in other domains that very different theories can sometimes explain the same results.[19] Third, by putting alternative theories into competition, we are forced to develop sharp tests—to search for properties of the data that can discriminate between the theories, in the sense of being explained by one of them but not the other. Insofar as one theory survives these additional tests, we have strengthened the arguments in favor of it. And fourth, by pitting alternative theories against the same set of data, and investigating the basis of their success or failure, we increase our knowledge of how different types of mechanisms behave.

Why should feature tests be carried out sequentially? For that matter, why should any two mental processes be carried out sequentially? In some cases, process P_2 might depend on information produced by process P_1. (For example, as suggested in figure 9.6, the feature-testing process, **FT**, depends on information provided by an encoding process, **E**, and the same-different decision, **D**, depends on information provided by **FT**.) The P_1-P_2 pair would then be described as data-dependent, as in the discussion of feature-testing and residual processes in section 9.2.3.1. But the individual feature tests for geometric forms (or letter tests for letter strings) that determine whether there is any mismatch are *not* data-dependent in this

sense. Another explanation for sequential structure is that the system that carries out P_1 and P_2 is inherently limited in *capacity*. Like the ordinary digital computer with a single central processing unit (cpu) that is capable of carrying out only one instruction at a time, this particular system can be expected to carry out only one test at a time.

An important alternative possibility is that capacity is not limited, and multiple feature tests can start simultaneously and be carried out in parallel. (By analogy, we can think of each feature test as a runner in a competition. As in the analogy, tests do not, in general, end simultaneously.) In an *unlimited-capacity* parallel testing process, the duration of each test is uninfluenced by the number of tests being carried out concurrently. (Think of each runner having a separate race track, and no information about the progress of the other runners.)

The general unlimited-capacity parallel-test mechanism we shall be considering has the following five properties:

1. Feature tests start simultaneously and are carried out in parallel;
2. The durations of tests that start together are mutually independent and are unaffected by the number of other tests that must be carried out;
3. No test is carried out more than once;
4. R_{diff} is initiated if and when a feature mismatch is discovered (the process is self-terminating); or
5. R_{same} is initiated if and when all n_{rel} tests are completed with no mismatch.

If there is more than one mismatch, $T_{\text{ft}}(\text{diff})$ is the duration of the fastest mismatching test, which is likely to decrease with the number of such tests. In contrast, if there are no mismatches, $T_{\text{ft}}(\text{same})$ is the duration of the slowest matching test, which is likely to increase with the number of such tests. These decreases and increases are called, respectively, "statistical facilitation" and "statistical inhibition" below.

9.2.5.1 Statistical Facilitation and the Effects of Process Variability

For a sequential model, the introduction of variability into the component operations without altering their means tends not to change qualitatively the pattern of $\overline{RT}$s that the model produces. Hence intuitions based on fixed-duration (deterministic) processes are usually helpful in thinking about variable-duration (stochastic) ones. In contrast, for parallel processes we can often be tricked by such intuitions. For example, suppose the parallel mismatching tests of features *A*, *S*, and *D* have durations γ_A, γ_S, and γ_D that have the same mean. If γ_A, γ_S, and γ_D were fixed from trial to trial (deterministic), we would expect no effect of n_{diff} on $\overline{RT}_{\text{diff}}$. On the other hand, suppose γ_A, γ_S, and γ_D varied independently from trial to trial.

Then, on average, as we increase n_{diff} from $n_{diff} = 1$ to $n_{diff} = 3$, and thus increase the number of concurrent tests that might eventuate in a mismatch and lead to the initiation of R_{diff}, the shorter would be the duration of the fastest of those tests, and hence the shorter $\overline{RT}_{diff}$. Because this effect looks like facilitation—acceleration of one or more component processes by an increase in n_{diff}—but is not, it is sometimes called "statistical facilitation" (Raab 1962).[20]

To see why statistical facilitation occurs, let us return to the racing analogy. Suppose a set of runners is selected for a one-mile race such that each runner has the same average time, but from race to race each runner fluctuates around the average, and suppose what is a good day for one runner might be a bad day for another. We might think of each runner as having a particular time for today, a time that could be revealed if the runner competed in today's race. Now suppose a random subset of the runners is selected to compete today. By making the subset bigger, the organizer increases the chance that the shortest time is very short. Thus the winning time will be shorter, on average, for a larger subset of runners. See comment 16 for a numerical example of statistical facilitation based on dice rolling, as well as an example of the complementary "statistical inhibition" property.

9.2.6 Parallel Tests: Effect of Number of Relevant Features on Mean Reaction-Time

Consider the effect of increasing the number of *relevant* features while keeping the number of mismatching features constant. For the mechanism with parallel tests described above, this should have no effect on $\overline{RT}_{diff}$, which depends on the fastest mismatching test. Because of the unlimited-capacity property, adding tests that would produce a match if permitted to go to completion can have no influence on any mismatches. In terms of the racing analogy, suppose runners who represent matches are found to be disqualified after the race, for example, by a blood test that reveals steroids. The winning time is then that of the fastest runner among the subset of the runners who represent mismatches. Adding runners who have no chance of winning has no influence on the winning time.

In contrast to this prediction, experiments have shown that $\overline{RT}_{diff}$ increases if n_{rel} is increased while n_{diff} is held constant, which argues against parallel models for those experiments, but is consistent with sequential testing. (See figure 9.4A for letter strings, and Nickerson 1972, figures 12 versus 13, for geometric patterns.)[21] How can the parallel mechanism be rescued from this difficulty, to accommodate the effect of n_{rel}? One approach would be to hypothesize that the effect is produced, not by the feature-testing stage (**FT** in figure 9.6) but by the encoding stage (**E**) that

precedes it. This stage presumably forms a representation of the second (test) stimulus to be used by the **FT** stage in comparing it to the representation already formed of the first stimulus. The duration of **E** could then increase with n_{rel}, but would be uninfluenced by n_{diff}. Where we have good information about the form of the n_{rel} effect (figure 9.7B), it appears strikingly linear (as expected from the sequential-test model); this tells us that any such effect of n_{rel} on T_e must be linear. In terms of equation 9.4 we would then have, for the *augmented parallel-test model*:

$$\overline{RT}_{diff} = \bar{T}_e(n_{rel}) + \bar{T}_{ft}(n_{diff}) + \bar{T}_d + \bar{T}_r, \tag{9.17}$$

where $\bar{T}_e(n_{rel})$ is of the form $A + Bn_{rel}$. One of the great attractions of the sequential-test model is its parsimony in being able to explain the full effects of both n_{rel} and n_{diff} on $\overline{RT}_{diff}$ in terms of a single process. Although it seems unfortunate to have to complicate matters, as the augmented model does, an effect of n_{rel} on **E** is not implausible.

On the other hand, if the effects of n_{diff} and n_{rel} on $\overline{RT}$ result from their influencing the durations of different stages of processing, as indicated in equation 9.17, it follows that the effects of each of these factors on $\overline{RT}$ must be invariant over levels of the other. For example, the amount by which $\overline{RT}_{diff}$ is reduced when n_{diff} is increased from 1 to 2 must be the same, $\bar{T}_{ft}(2) - \bar{T}_{ft}(3)$ regardless of whether $n_{rel} = 2$, 3, or 4. Conversely, the amount by which $\overline{RT}_{diff}$ is increased when n_{rel} is increased from 2 to 4, must be the same, $\bar{T}_e(4) - \bar{T}_e(2)$, regardless of whether $n_{diff} = 1$ or 2. Such invariance contrasts with the modulation of the effects of each factor by the other that is expressed in equation 9.8, and that appears to be confirmed by the data in figure 9.7B, for example. A corollary of the invariance property is the *additivity* of the effects of the two factors, n_{diff} and n_{rel} on $\overline{RT}_{diff}$: the combined effect of changes in both factors is the *sum* of their separate effects.[22]

It would be premature, however, to dismiss the parallel model on the grounds that additivity is violated in the data, without looking more closely at how well it fits and explicitly comparing it to the sequential model. How well can the $\overline{RT}_{diff}$ data be explained by a linear effect of n_{rel} that is additive with an effect of n_{diff}? As an alternative to equation 9.8, we thus need to consider how well the data can be fitted by equation 9.17, which is equivalent to

$$\overline{RT}_{diff} = bn_{rel} + g(n_{diff}), \tag{9.18}$$

where b is a constant, and $g(\ )$ a decreasing function. Because these quantities are unknown, we must use the data to estimate them, that is, we must fit b, $g(1)$, $g(2)$, $g(3)$, and $g(4)$ to the data. Fitting this model to the ten $\overline{RT}_{diff}$ data points thus requires us to estimate five "free parameters," as compared to the two parameters, α and θ of equation 9.8, estimated in

fitting the sequential-test model to the same ten data points. With two models that are equally valid or invalid, the one with more free parameters is likely to fit better because it can "capitalize on chance" more—that is, conform better to chance deviations in the data from the "true" values, due to sampling error. The augmented parallel model thus has a considerable advantage over the sequential.

Figure 9.8, A and B, shows the sequential and augmented-parallel models fitted to Bamber's $\overline{RT}_{diff}$ data (1969) in a way that makes it easy to note the magnitudes of the deviations of model from data and observe patterns in these deviations. For the augmented parallel model the estimated parameter values in milliseconds are $\hat{b} = 27.0$, $\widehat{g(1)} = 363.2$, $\widehat{g(2)} = 319.4$, $\widehat{g(3)} = 299.5$, and $\widehat{g(4)} = 282.2$. At first glance, the augmented parallel model fits well, but further inspection shows it to be inferior to the sequential model, despite the advantage conferred on it by its larger number of free parameters. The mean absolute deviation of model from data is 3.0 ms in panel A, and 4.1 ms in panel B.[23] The fitted values displayed in panel A show how n_{diff} modulates the effect of n_{rel} in the sequential-test model. The slope of the linear function relating $\overline{RT}$ to n_{rel} is reduced by each increase in n_{diff}, and this change in slope describes the data quite well. This effect in the data is also shown by their deviations from the parallel lines of panel B, lines that reflect the additivity required by the parallel-test model. I have already commented (section 9.2.4.1) on the inequality of the values of the four means for which $n_{diff} = n_{rel}$, which increase as n_{diff} increases from 2 to 4. As shown in panel B, a model in which n_{rel} and n_{diff} influence different stages can accommodate such an effect; indeed, the fitted effect is larger than what is seen in the data. Also evident in panel A is the deviation noted in section 9.2.4.1 between the height of the linear function for $n_{diff} = 2$ and the height of the corresponding data points.[24]

Based on the data and the analysis thus far of the properties of the two contending models, we should favor the sequential-test model. However, the augmented parallel model is worth further investigation, for four reasons. First, the sequential model has some defects. Second, the parallel model does provide an approximate fit. Third, arguments from the R_{same} data might increase the credibility of the parallel model. And fourth, further consideration of the parallel model might provide insights about how parallel processes behave and how models can be tested, which are among the goals of the present chapter.

9.2.7 Parallel Tests: Effect of Number of Mismatching Features on Mean Reaction-Time

We can better come to understand the properties of $\overline{RT}_{diff}$ for a parallel mechanism if we impose additional constraints. The separate-racetracks

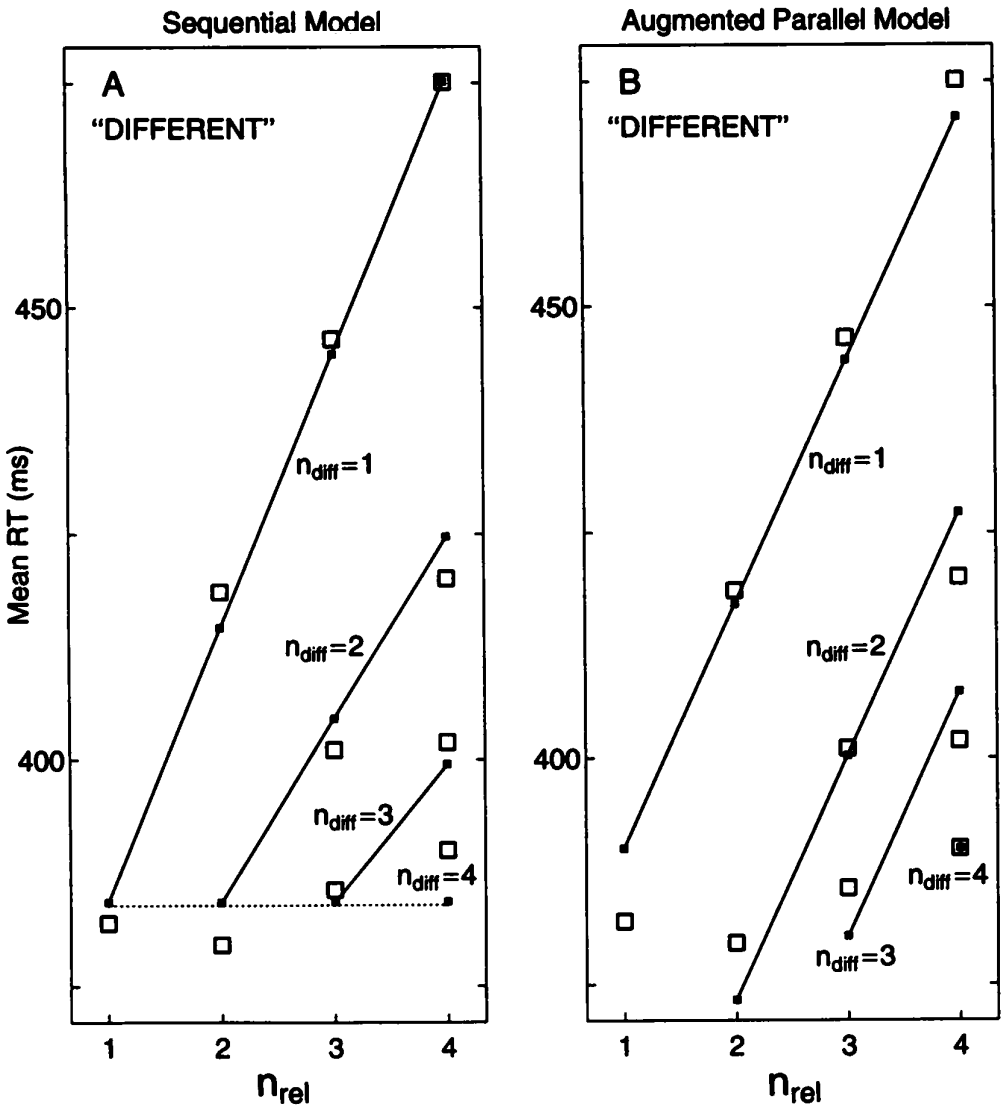

Figure 9.8
Two models fitted to Bamber's $\overline{RT}_{diff}$ data (1969). Data points are shown as open squares; fitted model values as small solid squares connected by lines. This plotting method makes it easy to appreciate deviations of model from data. In panel A, $\overline{RT}_{diff}$ values are fitted by the sequential-test model (equation 9.8, two fitted parameters), also shown in figure 9.7A. In panel B, $\overline{RT}_{diff}$ values are fitted by the augmented parallel-test model (equation 9.18, five fitted parameters, effect of n_{diff} additive with effect of n_{rel}, which is constrained to be linear). In panel C, $\overline{RT}_{same}$ values are fitted by two versions of the augmented parallel-test model with the effect of n_{rel} estimated from the $\overline{RT}_{diff}$ data. The broken line corresponds to a nil effect of n_{rel} on $\overline{T}_{ft}$; the solid curve corresponds to an effect generated as described in the text. The heights of both fitted functions were determined by least squares fitting.

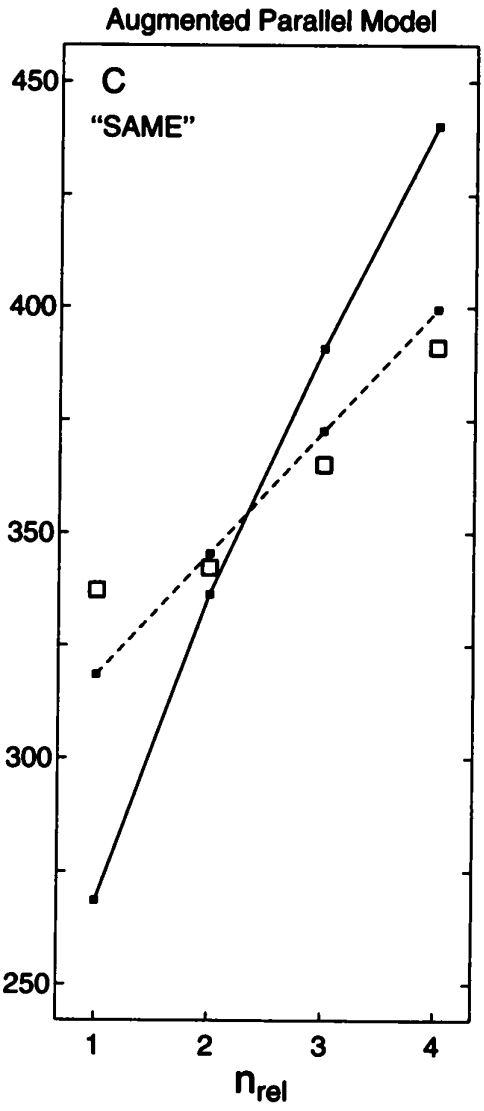

Figure 9.8 (cont.)

analogy (section 9.2.5) helps make clear that in an unlimited-capacity parallel process the number or durations of any matching tests have no influence on the time to complete the first mismatching test. That is, given any particular $n_{diff} \geq 1$, n_{rel} ($n_{rel} \geq n_{diff}$) has no influence on the $\overline{RT}_{diff}$ for any such mechanism. Because our concern in the present section is limited to R_{diff}, the additional constraints need therefore apply only to the durations of mismatching feature tests. Imposition of the constraints leads to variants, or special cases, of the general parallel mechanism. In this section I define four such variants and consider what each of these models implies about the effect of n_{diff} on $\overline{T}_{ft}$, and hence on $\overline{RT}$. The variants differ with respect to the equality across attributes (or letter positions) of mismatching test durations, and with respect to the variability of these durations

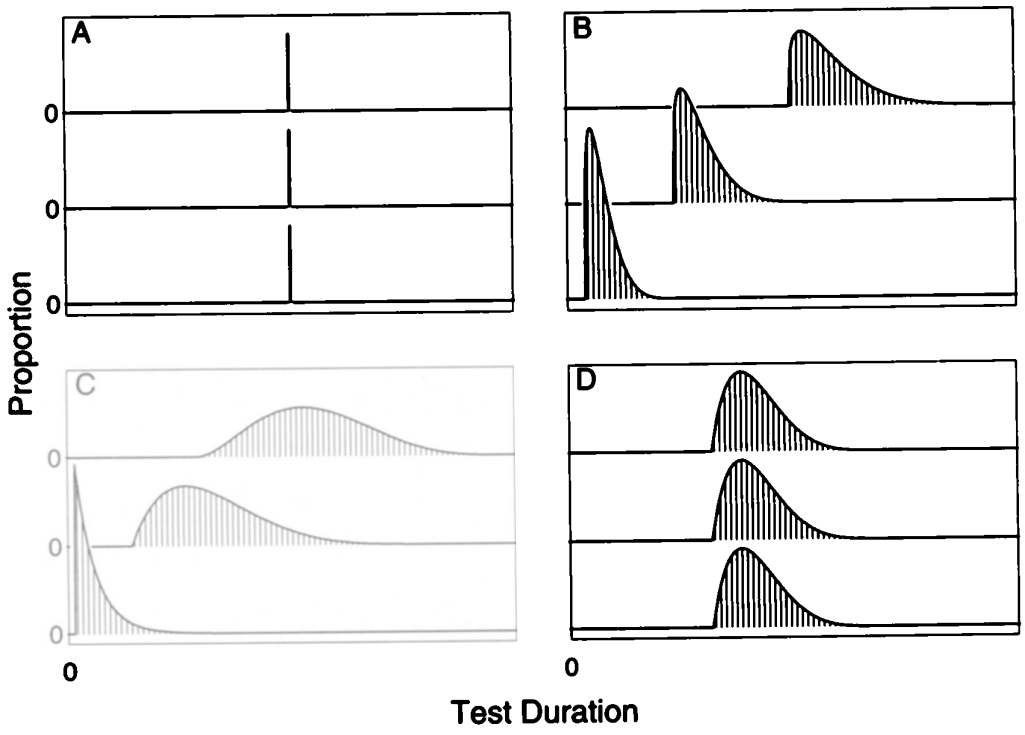

Figure 9.9
Hypothetical distributions of mismatching test durations for three hypothetical features in four variants of parallel-test mechanism. The four panels correspond to the four variants, and the three distributions in each panel correspond to the three features. Duration is represented on the x-axis, and (roughly speaking) the y-value for a duration, x, represents the proportion of durations that take on the value x. The range of possible durations for a given test is reflected by the interval on the x-axis over which the proportion is nonzero. In panel A, mean test-durations for all three features are equal and there is no variability (variant 1). In panel B, test durations have different means and are variable, but the variability is small enough relative to the differences among the means that duration ranges for different features do not overlap (variant 2). Also, these three distributions have the same shape, but differ in mean and spread (see section 9.2.7.4). In panel C, means differ but there is enough variability relative to the differences that ranges overlap (variant 3). Also, the bottom example is an exponential distribution (see section 9.2.7.4). In panel D, means are equal and the three distributions are identical; thus duration ranges overlap perfectly (variant 4).

from trial to trial. By considering such variants, even if some are unrealistic, we can usefully educate our intuition.

9.2.7.1 Parallel Variant 1: Equal Fixed Test-Durations

In this variant, all mismatching tests are assumed to have the same fixed duration, as illustrated in figure 9.9A. Given this constraint, there can be no effect of $n_{diff} \geq 1$. Measuring from the start of the testing process, all mismatching tests would be completed after the same elapsed time, as if all the runners in a competition crossed their finish lines simultaneously; an increase in the number of mismatches—or the number of runners—

would therefore confer no time advantage. Because an increase in n_{diff} clearly reduces $\overline{RT}_{diff}$ we have evidence against this variant.

> *Comment 12: Limited-capacity parallel testing.* Another possibility, in addition to sequential tests and unlimited-capacity parallel tests, and in the spirit of variant 1, is a *limited-capacity* parallel process. Suppose a fixed amount, C, of capacity is allocated equally among a set of n_{rel} ongoing processes, so the ith process has capacity $c_i = C/n_{rel}$. Suppose further that the rate r_i of the ith process is a function, $r_i(c_i)$, of the capacity allocated to it. By choosing an appropriate function we can explain any form of increase of $\overline{RT}$ with number of tests. (Application of such a model to the n_{rel} effect is therefore an instance of accommodating rather than constraining the data, discussed in comment 14, section 9.2.7.4. With this much flexibility, we need other evidence to justify a particular choice of rate-capacity function. One such justification would be the finding of the same function in diverse experiments.) As an example of such a function, suppose that r_i is proportional to the capacity allocated to it. If the processes have the same proportionality constant in their rate-capacity relationship, then r_i is inversely proportional to n_{rel}. Because the duration of a process is inversely proportional to its rate, it follows that the time taken by the ith test is directly proportional to n_{rel}. The n_{rel} processes will therefore end simultaneously, after an elapsed time proportional to n_{rel}. A parallel mechanism can thus produce sequential-looking data. Like variant 1, however, this arrangement does not respond appropriately to changes in n_{diff}.

9.2.7.2 Parallel Variant 2: Unequal Mean Test-Durations with Limited Variability

In this variant, the mean test-durations for different attributes (or letter positions) can differ, and for a particular attribute the time to discover a mismatch can vary from trial to trial. However, the trial-to-trial duration variability for a test of any particular attribute is sufficiently small, relative to the duration differences across attributes, that the ranges of test durations for different attributes do not overlap (see figure 9.9B). For example, with two relevant features, the shortest duration of the slower of the two tests would exceed the longest duration of the faster. In terms of the racing analogy, no matter how many races were run by these competitors, the runner whose average time was shortest would win every race and would never have a day sufficiently bad to be beaten by even the usual runner-up. Suppose a mismatch of A generates R_{diff} faster than a mismatch of any other single attribute. If so, adding more mismatches to an A-mismatch cannot speed the response, given this variant. That is, given

an A-mismatch there should be no effect of introducing additional mismatching features. (If A-mismatch is the reigning champion runner, then A-mismatch will win, regardless of the number of competing runners.) Because data (Nickerson 1972, figures 11 and 12) show that an increase in $n_{diff} \geq 1$ reduces $\overline{RT}_{diff}$, even when the condition with $n_{diff} = 1$ is created by selecting the attribute for which single mismatches are detected fastest, we can reject parallel variant 2. (However, this finding is consistent with sequential models if there are any trials on which the most discriminable—that is, fastest—attribute is not the first to be tested.)

> *Comment 13: Zero variability.* In my view, an assumption of zero variability of the duration of any biologically controlled process, as in parallel variant 1, is highly suspect. Sometimes, however, a prediction from an extreme assumption such as this one holds, even when the assumption is relaxed, as long as it is not relaxed too much. Thus, because the duration variability is limited in variant 2, the statistical facilitation described in section 9.2.5.1 is absent, just as if the duration variability were zero.

9.2.7.3 Parallel Variant 3: Variable Test-Durations with Unconstrained Means

In this variant, mismatch durations for the same feature (or letter position) vary from trial to trial, generating a range of such durations, and this variation is great enough relative to differences among the means so there may be overlap of the duration ranges for different features (see figure 9.9C). We can again think of all the mismatching feature tests on a trial as runners in a competition, with the winner generating the earliest mismatch and initiating R_{diff}. Suppose feature mismatches for A are, on average, faster than those for S. If their ranges overlap, then on some trials S will happen to be faster than A and win the race (another form of statistical facilitation; section 9.2.5.1), so that supplementing an A mismatch with an S mismatch will, on average, shorten $\overline{RT}_{diff}$, even though S is slower than A, on average.

A simplified numerical example of this phenomenon is given in table 9.5. Consider the third column, which contains the durations of mismatching tests. In the first three pairs of rows ($n_{diff} = 1$), we see the possible test durations when the two patterns differ by just one feature—A, S, or D. Consider the first pair of rows, where A is the mismatching feature. As indicated, suppose the duration of a mismatching test of A is equally likely to be either $\gamma_A = 100$ ms or $\gamma_A = 200$ ms. For the second pair of rows, where only the S feature differs, the duration of the mismatching test of S is equally likely to be $\gamma_S = 150$ ms or $\gamma_S = 250$ ms. (I am not seriously proposing that two-point distributions, where all possible durations are concentrated at two values, are more plausible in this case than

Table 9.5
Effect of n_{diff} ($1 \leq n_{diff} \leq 3$) on $\overline{RT}_{diff}$ in a parallel-testing process (variant 3) with overlapping mismatch-duration distributions (durations in ms)

n_{diff}	Mismatching feature(s)	Test durations	Proportion	Shortest duration	Mean shortest	Grand mean
1	A	100	.50	100	150	200
		200	.50	200		
1	S	150	.50	150	200	
		250	.50	250		
1	D	200	.50	200	250	
		300	.50	300		
2	A,S	100, 150	.25	100		154
		100, 250	.25	100		
		200, 150	.25	150		
		200, 250	.25	200		
2	A,D	100, 200	.25	100		
		100, 300	.25	100		
		200, 200	.25	200		
		200, 300	.25	200		
2	S,D	150, 200	.25	150	175	
		150, 300	.25	150		
		250, 200	.25	200		
		250, 300	.25	250		
3	A,S,D	100, 150, 200	.125	100	138	138
		100, 150, 300	.125	100		
		100, 250, 200	.125	100		
		100, 250, 300	.125	100		
		200, 150, 200	.125	150		
		200, 150, 300	.125	150		
		200, 250, 200	.125	200		
		200, 250, 300	.125	200		

are continuous distributions, where any value within some range is possible, but they are useful for illustration.) The means for A and S alone are then $\bar{\gamma}_A = 150$ and $\bar{\gamma}_S = 200$ ms, respectively, so that γ_A is the smaller, on average. The next three sets of four rows list possible pairs of mismatching test durations on trials where $n_{diff} = 2$; two features—A and S, for example—both differ. Note that although γ_A is smaller, on average, than γ_S, the distributions overlap, so the combination $\gamma_A = 200 > \gamma_S = 150$ ms (row 9) is possible.[25] The final set of eight rows list possible triples of test durations on trials on which $n_{diff} = 3$; all three features differ. For example, the first of these last eight rows indicates that on 1/8 of such trials, the test durations for A, S, and D would be 100, 150, and 200 ms, respectively.

The last column of the table illustrates that the average time required to achieve a mismatch under variant 3 decreases as the number of mismatching features increases, just as in the model with sequential feature tests; this finding therefore cannot, by itself, discriminate between the sequential and parallel models. Moreover, the decline in average time as n_{diff} increases shows "diminishing returns," just as for sequential tests: the reduction from $n_{diff} = 1$ to $n_{diff} = 2$ (which is $200 - 154 = 46$ ms) is greater than the reduction for $n_{diff} = 2$ to $n_{diff} = 3$ (which is $154 - 138 = 16$ ms). The function relating the average time to n_{diff} is thus concave up.[26]

The assumption incorporated in table 9.5 is that the durations (γ_S and γ_A, for example) of mismatching tests that start together are *independent*. This assumption permits us to assert that the possible pairs (or triples) of values are equally probable. Assumptions of independence are very convenient, and for some models may be necessary, to easily derive predictions, but they are not necessarily plausible. For example, if there were trial-to-trial variation in the overall processing effort, then on some trials all test durations might tend to be longer than on other trials, which would cause a violation of independence and generate a positive correlation of test durations. To see the importance of the assumption, let us consider two extreme alternatives for trials on which A and S both mismatch (rows 7–10 of table 9.5). First, suppose the A and S mismatch durations have a strong positive correlation: either both take on their low values, $\gamma_A = 100$ and $\gamma_S = 150$, generating 100 ms as the shortest duration, or both take on their high values, $\gamma_A = 200$ and $\gamma_S = 250$, generating 200 ms as the shortest duration. The resulting mean is 150 ms, greater than the 138 ms in table 9.5, and equal to the value for A alone (rows 1–2). There is no statistical facilitation in this case. Second, suppose the A and S mismatch durations have a strong negative correlation: either $\gamma_A = 100$ and $\gamma_S = 250$, generating 100 ms as the shortest duration, or $\gamma_A = 200$ and $\gamma_S = 150$, generating 150 ms as the shortest duration. This maximizes the amount of statistical facilitation. The resulting mean is 125 ms, less than the 138 ms in table 9.5. These examples reveal an important respect in which parallel

and sequential models differ: Whereas predictions of $\overline{RT}$ based on sequential models are unaffected by the correlations of test durations, predictions of $\overline{RT}$ based on parallel models can be sensitive to them; also, such predictions, again unlike those of sequential models, depend on what is assumed about the form of the distributions of test durations.

In contrast, one of the properties that makes sequential models pleasantly tractable is that predictions about means do not depend on the independence of test durations or the forms of their distributions. The following numerical example, related to the one in table 9.5, should help to make this clear. Suppose we are interested in $\overline{RT}_{same}$ for a sequential process, that the mismatching features in the table are instead matching features ($n_{diff} = 0$), and that $n_{rel} = 2$. Then the relevant quantity is the sum of the test durations rather than the shortest duration. The four sums associated with the A, S section of table 9.5 are 250, 350, 350, and 450 ms, respectively, whose mean is 350 ms. Now suppose we have the extreme positive correlation mentioned above. The two sums are 250 and 450 ms, with the same mean. And the two sums associated with the extreme negative correlation are both 350 ms, again with the same mean.[27]

9.2.7.4 Parallel Variant 4: Variable Test-Durations with Equal Means and Identical Distributions

Here the duration of the test of a particular feature or element varies from trial to trial, making this variant more plausible than variant 1, but unlike variant 2, all features or elements are identical, in the sense that their test durations are indistinguishable (see figure 9.9D). While this property is unlikely to hold for geometric patterns that have not been carefully adjusted, it might apply to letter-string patterns, as mentioned in section 9.2.4.5. Like variant 3, the duration ranges of different tests overlap, and because of the resulting statistical facilitation, an increase in n_{diff} produces a reduction in $\overline{RT}_{diff}$, an effect that is qualitatively similar to that obtained from self-terminating sequential tests.

Can we say anything more precise about this expected effect of n_{diff}? The effect can be described in terms of two characteristics, its *shape* and its *size*. By "shape" I mean the *relative* sizes of the one-step reductions in $\overline{T}_{ft}$ caused by increasing n_{diff} from 1 to 2, from 2 to 3, and from 3 to 4. Consider the data for $n_{rel} = 4$. If $\overline{RT}_{jk}$ is the mean RT in ms when $n_{rel} = j$ and $n_{diff} = k$, we have $\overline{RT}_{41} = 475.1$, $\overline{RT}_{42} = 420.1$, $\overline{RT}_{43} = 401.9$, and $\overline{RT}_{44} = 390.1$. The one-step reductions are then $\overline{RT}_{41} - \overline{RT}_{42} = 55.0$, $\overline{RT}_{42} - \overline{RT}_{43} = 18.2$, and $\overline{RT}_{43} - \overline{RT}_{44} = 11.8$ ms. The shape of the effect can then be obtained by dividing each of these differences by the first, which gives the three values 1.00, 0.33, and 0.21.[28] (The first value is 1.00 by definition, of course, but is included for clarity.) In the decline of these

values we again see the diminishing returns of adding a mismatch. For a given effect shape, the first of the differences, $\overline{RT}_{41} - \overline{RT}_{42}$, provides a measure of the effect size.[29] To explain how the shape and size of the n_{diff} effect are predicted by the parallel-test model, let us turn briefly to the distributions of test durations, such as those illustrated in figure 9.9, and consider their shapes, locations, and spreads.

Consider the duration γ of a mismatching test. The *distribution* of γ describes how γ varies over a large number of such tests. Like the hypothetical distributions illustrated in figure 9.9, a distribution describes the set of values that γ can assume (all those values—on the x-axis—whose proportion—on the y-axis—is nonzero) as well as the proportion of occurrences of each value. Distributions may differ in *shape*. For example, if the distribution is "positively skewed," then γ is small most of the time, with occasional large values (like most of the distributions in figure 9.9). The peak of such a distribution is toward the left, and the long tail is on the right. (The three distributions in figure 9.9C increase in positive skewness from top to bottom.) Distributions of the same shape can have different *locations*, associated with translations along the x-axis. (A translation of c ms to the right, for example, increases the mean and median of the distribution by c ms.) Distributions of the same shape can also have different *spreads*, associated with scaling (multiplication) of the x-axis. The standard deviation and the variance are particular measures of spread.[30] (Increasing the spread by a factor k, for example, increases the standard deviation of the distribution by that same factor, and increases the variance by the factor k^2.) The three distributions in figure 9.9B have the same shape, but differ in mean and spread.

The shape of the n_{diff} effect is determined by the shape of the distribution of γ. Simulations with several of the common distributions indicate that an effect shape close to the one observed in Bamber's experiment can be achieved and, further, that its high degree of diminishing returns requires a distribution that is strongly positively skewed, such as the *exponential distribution*, illustrated by the bottom distribution in figure 9.9C.[31] (The effect shape produced by simulations with the exponential distribution is 1.00, 0.33, 0.17, which agrees fairly well with the shape observed; the effect of adding a second racer is six times as great as the effect of adding a fourth. As an example of a contrasting case, the effect shape produced by simulations with the rectangular distribution, which is symmetric rather than skewed, is 1.00, 0.52, 0.35; the effect of adding a second racer is only three times as great as the effect of adding a fourth.)[32]

Whereas the shape of the n_{diff}-effect is determined by the shape of the distribution of γ, the size of the effect is determined by the distribution's spread, which can be measured by its standard deviation, sdev(γ). If we had confidence in our choice of shape (exponential), for the distribution,

we could then use the observed size of the n_{diff} effect (55.0 ms) to "predict" what sdev(γ) must be. (The simulations show that an n_{diff} effect of the desired size is produced if γ has an exponential distribution with sdev(γ) = 81 ms.) How can we use such a prediction of sdev(γ) in evaluating the parallel-test model? If we could directly measure sdev(γ), we could test the model by comparing this measurement to the prediction. When $n_{diff} = 1$, the duration, T_{ft} of the testing process (**FT** in figure 9.6) is the same as the duration, γ, of a single mismatching test. Thus what we need for comparision to the prediction is sdev(T_{ft}) when $n_{diff} = 1$, which we can write sdev($T_{ft}|n_{diff} = 1$). We cannot measure T_{ft} alone, however, but only when combined with the durations of the three other stages, as shown in equation 9.4. However, it is reasonable to believe that by concatenating other operations with **FT**, each of whose duration is likely to be variable, we can only increase spread. The observed RT spread is likely to be at least as great as the T_{ft} spread: sdev($T_{ft}|n_{diff} = 1$) $\leq$ sdev($RT|n_{diff} = 1$). Thus, if the sdev(RT) observed when $n_{diff} = 1$ was smaller than the sdev(T_{ft}) required by the size of the n_{diff} effect, we would have evidence against the parallel-test model. However, this conclusion depends on our being confident of our decision about the shape of the distribution of the mismatching-test duration, γ, confidence that is hard to justify.

Happily, a version of this argument about the size of the n_{diff} effect is available that does not depend on the particular form of the distribution of γ. Regardless of the shape of the distribution, within a set that includes all plausible such shapes, there is an upper bound on the amount of statistical facilitation, a bound that depends solely on the spread of the distribution (David 1970, equation 4.2.6). Let $\min_n(\gamma)$ be the smallest value in a sample of n independent values of γ. As n grows, the average value $\overline{\min_n(\gamma)}$ of $\min_n(\gamma)$ shrinks. The amount of such statistical facilitation is the extent to which $\overline{\min_n(\gamma)}$ is less than the mean test-duration $\bar{\gamma}$. This difference has an upper bound that is proportional to the spread:

$$\bar{\gamma} - \overline{\min_n(\gamma)} \leq \frac{(n-1)}{\sqrt{2n-1}}\text{sdev}(\gamma). \tag{9.19}$$

It follows that the amount of statistical facilitation determines a *lower bound* on sdev(γ):

$$\text{sdev}(\gamma) \geq \frac{\sqrt{2n-1}}{(n-1)}\{\bar{\gamma} - \overline{\min_n(\gamma)}\}. \tag{9.20}$$

For the reasons given above, when RTs are collected under conditions in which $n_{diff} = 1$, sdev(γ) in equation 9.20 can be replaced by sdev($RT|n_{diff} = 1$). For $n = 2$ which we have, in Bamber's experiment, when $n_{diff} = 2$, the factor in braces on the right is estimated by $\overline{RT}_{41} - \overline{RT}_{42}$, and the

inequality becomes $\text{sdev}(RT|n_{\text{diff}} = 1) \geq \sqrt{3}\ (\overline{RT}_{41} - \overline{RT}_{42}) = 95$ ms. If the observed value of $\text{sdev}(RT|n_{\text{diff}} = 1)$ proved to be less than 95 ms, this would indicate that the n_{diff} effect in this experiment is too great to be explained by statistical facilitation, and we would have evidence against variant 4 of the parallel-test model. Unfortunately, because the RT variability measures from Bamber's experiment are no longer available, this test awaits further data collection.

> *Comment 14: Constraining versus accommodating the data.* As we have seen, the magnitude and form of the effect of n_{diff} produced by a parallel mechanism depends on many details of the mechanism, such as the variability of the test duration, the shapes of the duration distributions, and the correlations among durations. In contrast, for a sequential model that includes constraint 4 (tests of different attributes have the same mean duration), the magnitude of the n_{diff} effect (for given n_{rel}) depends only on β, and the form of the effect is fixed, regardless of any of these properties of the test durations. Hence, the relation between $\overline{RT}_{\text{diff}}$ and n_{diff} can falsify all members of the class of models that describe sequential tests (with constraint 4), but a large variety of such relations can be accommodated by members of the class of models with parallel tests. Conversely, if the n_{diff} effect is well described by a sequential model, as in the data of figure 9.7A, this provides stronger support for sequential tests than for tests in parallel because only those models in a relatively small subset of possible parallel models are consistent with such a pattern. In general, the larger the variety of mutually incompatible data patterns consistent with a theory (i.e., the more flexible or "weaker" the theory), the less persuasive is any one of those patterns as evidence that supports that theory. Howson (1990, 226) puts it this way: "Of two rival theories, initially equally well supported, but differing in that one independently predicts data that the other merely absorbs into the evaluation of a free parameter, the former receives the greater support from those data" (see also Howson and Urbach 1993).

9.2.7.5 Status of the Parallel-Test Model

A parallel mechanism with unlimited capacity cannot account for any effect of n_{rel} on $\overline{RT}_{\text{diff}}$, and must therefore be augmented by another mechanism to do so; one plausible possibility is an encoding stage (**E**) that precedes the feature-testing process **FT**, and whose duration increases appropriately with the number of characters (relevant features for geometric stimuli). Thus the attractive parsimony of the sequential-test model must be sacrificed if the parallel model is to accommodate the data. One consequence of having separate stages in which the n_{rel} and n_{diff} effects

operate is that they should be additive. Bamber's data depart systematically from such additivity, in the direction expected for the sequential-test model, although the departure is not very great; more data are needed.

What about the effect of n_{diff} on $\overline{RT}_{diff}$? For parallel mechanisms with independent test durations, either n_{diff} has no effect on $\overline{RT}_{diff}$ (variants 1 and 2), or $\overline{RT}_{diff}$ declines as n_{diff} increases (variants 3 and 4) because of statistical facilitation. Thus a decline of $\overline{RT}_{diff}$ with n_{diff}, which is typically observed, does not preclude parallel tests. Furthermore, simulations suggest that a distribution of γ can be found that can explain the shape of the n_{diff} effect. However, the size of the n_{diff} effect requires the spread of this distribution to be relatively large; it has yet to be determined whether the data are consistent with this requirement.

9.2.8 Sequential versus Parallel Tests: Inferences Based on Differential Mismatch-Durations

We have seen examples of model properties that depend on the similarity of the durations of mismatching tests for different attributes. For sequential tests, for example, constraint 4 (equal duration means) turned out to be critical for developing certain predictions (section 9.2.4.5), and for parallel tests, identical duration distributions permitted interesting inferences (section 9.2.7.4). Model tests that exploit heterogeneity among mismatch durations are also of interest, especially as heterogeneity may be easier to achieve experimentally.

Let F and S denote features for which mismatching tests are fast and slow, respectively, so that $\bar{\gamma}_F < \bar{\gamma}_S$. (In some studies, color and shape would qualify as F and S, respectively.) It will be convenient to describe the relationship between the stimuli compared on a trial as I have in table 9.1: features that are relevant and that match are in lowercase italics (f, s), and features that are relevant and mismatch are in uppercase bold (**F**, **S**). The required relation between mismatch times could be established by showing that $\overline{RT}(\mathbf{F}) < \overline{RT}(\mathbf{S})$, whether tests are sequential or parallel.

Consider what happens to the duration T_{ft} of stage **FT** (figure 9.6) when we add a slowly tested mismatching feature to a rapidly tested one. What is the relation between $T_{ft}(\mathbf{F})$ and $T_{ft}(\mathbf{F},\mathbf{S})$?[33]

For parallel variants 3 and 4, the distributions of γ_F and γ_S overlap, so that on some (small) proportion of the trials the test of S will be completed before the test of F, yielding $T_{ft}(\mathbf{F},\mathbf{S}) < T_{ft}(\mathbf{F})$. (It may be surprising that adding a slow feature to a fast one can *reduce* the mean test duration.) For parallel variants 1 and 2, the test of F will always be completed first, yielding $T_{ft}(\mathbf{F},\mathbf{S}) = T_{ft}(\mathbf{F})$. For parallel tests in general, then, we expect

$$T_{ft}(\mathbf{F},\mathbf{S}) \leq T_{ft}(\mathbf{F}). \tag{9.21}$$

For sequential tests, what happens when **S** is added to **F** depends on the search path. If *F* is always tested before *S* then there should be no effect, and we expect $T_{ft}(\mathbf{F},\mathbf{S}) = T_{ft}(\mathbf{F})$. If *S* is tested first on even a small proportion of the trials, we expect $T_{ft}(\mathbf{F},\mathbf{S}) > T_{ft}(\mathbf{F})$. For sequential tests in general, then, we expect

$$T_{ft}(\mathbf{F},\mathbf{S}) \geq T_{ft}(\mathbf{F}). \tag{9.22}$$

Because inequalities 9.21 and 9.22 both permit the two terms to be equal, observing *equality* would not allow us to distinguish between sequential and parallel mechanisms. If we observe *inequality*, however, we may have a discriminating test. Happily for deciding between the mechanisms, what is observed for geometric forms when single-feature $\overline{RT}$s are known to differ is an inequality consistent with equation 9.22: $\overline{RT}(\mathbf{F},\mathbf{S}) > \overline{RT}(\mathbf{F})$. (See Hawkins 1969, table 2; or Nickerson 1972, figure 13.) This finding has been used to argue for sequential and against parallel tests. If you favor a sequential-test model applied to Bamber's experiment (1969), you might argue that it is because mean mismatch durations for different letter locations differ minimally that approximate equality is found there (filled points in figure 9.7B), and you might claim that, if anything, the values increase with $n_{diff} = n_{rel}$, consistent with equation 9.22. Unfortunately, the argument favoring sequential tests depends on a questionable assumption. (Before reading further, consider what this might be.)

How could the parallel-test model accommodate the finding that adding a slowly tested mismatching feature increases $\overline{RT}$? The argument above depends on the assumption that the only duration that is altered by adding the feature is T_{ft}. Given parallel tests, T_{ft} cannot be prolonged by adding a mismatching feature. But suppose another of the processes contributing to the *RT* (figure 9.6) is prolonged. (We have already seen, in section 9.2.6, that for a parallel mechanism to explain the effect of n_{rel}, we have to assume that an increase in n_{rel} causes an increase in the duration of a process other than **FT**, and I suggested the encoding stage **E** that precedes it as a likely possibility.) If T_e were prolonged by the added feature more than T_{ft} were shortened, it would then be possible for $\overline{RT}(\mathbf{F},\mathbf{S}) > \overline{RT}(\mathbf{F})$, at the same time as $T_{ft}(\mathbf{F},\mathbf{S}) \leq T_{ft}(\mathbf{F})$.

Is there an alternative manipulation that would also exploit the special properties of parallel tests, but would avoid influencing the durations of two different stages? Suppose we kept the number of relevant features constant: instead of adding a slowly tested mismatching feature, we changed a slowly tested feature from match to mismatch. Instead of equation 9.21, we would then have $T_{ft}(\mathbf{F},\mathbf{S}) \leq T_{ft}(\mathbf{F},s)$, and the encoding stage would have the same duration in both of these conditions. Unfortunately, however, this is nothing other than the n_{diff} effect, which, qualitatively, can

also be explained by sequential tests. Without predictions that are more quantitative, this effect cannot help us to discriminate between models.

9.2.9 Sequential versus Parallel Tests: Conclusions from "Different" Responses

We have seen that the sequential-test model accounts for details of the R_{diff} data with considerable parsimony. We have also seen that by adding a few "bells and whistles" the parallel-test model can also be made to do a fairly good job, but is still inferior. What might we do to sharpen the discrimination between these models?

First, it would be helpful to have more data of the same kind from an experiment like Bamber's (1969), to determine which of the deviations from each of the two models are reliable. Second, from the same sort of experiment, it would be helpful to have information about RT variability and how it is influenced by n_{rel} and n_{diff}. We have already seen, in section 9.2.7.4, how such information can be used to assess the parallel-test model. Also, if we strengthen the sequential-test model by incorporating an additional defining property, the derived properties ("predictions") can be extended to include an explicit statement about how the variance of the RT should be influenced by n_{rel} and n_{diff} that could be compared to the data, just as equation 9.8 makes such a statement about the behavior of $\overline{RT}$. (The additional defining property is that test durations as well as the durations of the other operations diagrammed in figure 9.6 are *stochastically independent*; this is roughly equivalent to there being no correlations among these durations. If such a strengthened model is supported, then the weaker model, without the added property, inherits the support. Of course, if we failed to find the predicted behavior of the variance, we might not be able to decide whether this was because the added property did not apply, or because the basic model was at fault.)

A third approach would be to conduct variants of Bamber's experiment designed to be sensitive to important differences between the parallel and sequential models. For example, in a sequential mechanism, the tests on each trial must be carried out in some order. That is, there is a search path defined on each trial. Given the search path, and the self-termination property, RT_{diff} must vary systematically with the location of the first mismatching feature within that path: that is, $\overline{RT}$ should increase systematically (and possibly linearly) with the *serial position* within the search path of the first mismatch. If as experimenters we could gain some control over the ostensible search path, then this expectation could be tested. The challenge would be to achieve such control without changing the testing mechanism, and to demonstrate that we have done so.[34] If the underlying mechanism was one of independent parallel tests, then either there should

be no response to the experimental manipulation, or the mechanism should change to one that can respond.

> *Comment 15: Problem of multiple strategies.* Many psychologists believe that people are flexible, in the sense that they can choose among different combinations of mental operations to perform the same mental task. Because of this hypothesized freedom of choice, the combinations are called "strategies." (For example, suppose sequential and parallel tests are alternative strategies.) Such strategies and their choice need not be deliberate or conscious. Ideally, the selection of strategies should be brought under experimental control, rather than being left to the subject; the psychologist's goal of describing a strategy in detail should be separated from the study of what governs the selection of strategies. Unfortunately, until at least one of the strategies that people use in a task is described in detail, it is hard to determine whether an experimenter is investigating a pure strategy or a mixture of two or more, mixed from trial to trial or from subject to subject. One reason to prefer experimental paradigms whose data can be described by simple theories is that such parsimony may reflect a control over strategy achieved by those paradigms.

In the absence of such additional information, we are led to favor the parsimonious account of the $\overline{RT}_{\text{diff}}$ data provided by the sequential-test model. Unfortunately, as mentioned in section 9.1.4, this elegant account cannot also explain the $\overline{RT}_{\text{same}}$ data.

9.3 Reaction Time to Judge "Same"

9.3.1 Difficulties for Sequential Tests

In the discussion above, some of the properties of $\overline{RT}_{\text{same}}$ have already been developed (see equations 9.3, 9.7, 9.9, and 9.12). Here we consider four issues: (1) the relation between the magnitudes of $\overline{RT}_{\text{same}}$ and $\overline{RT}_{\text{diff}}$, (2) the rate at which $\overline{RT}_{\text{same}}$ increases with n_{rel}, (3) the linearity of the function relating $\overline{RT}_{\text{same}}$ to n_{rel}, and (4) the effect on $\overline{RT}_{\text{same}}$ of adding relevant matching attributes that are more discriminable.

1. *Speed of "same" versus "different".* As we have already seen (figure 9.7A), a mechanism with sequential tests, together with the four constraints of section 9.2.3.2, lead us to expect that for $n_{\text{rel}} > 1$, $\overline{RT}_{\text{same}} > \overline{RT}_{\text{diff}}$. Yet the observed inequality is reversed in the geometric-pattern data with $n_{\text{rel}} = 3$ (figure 9.3A) and the letter-string data with $n_{\text{rel}} = 2$, 3, and 4 (figure 9.4A). We have seen, however, that by relaxing two of the constraints, so we permit $\alpha_{\text{same}} < \alpha_{\text{diff}}$ (section 9.2.4.3) and perhaps also

$\beta < \gamma$ (section 9.2.4.4), we can account for the "fast same" phenomenon. (An example that shows the sequential model fitting the average difference between $\overline{RT}_{same}$ and $\overline{RT}_{diff}$ is given in figure 9.7C by the dotted and dashed parallel lines.)

2. *Rate of increase of $\overline{RT}_{same}$ with number of relevant features.* Recall that, in the sequential mechanism, R_{diff} is preceded by from 0 to $n_{rel} - 1$ matches followed by one mismatch, and that R_{same} is preceded by n_{rel} matches. It follows that for both responses, the sequential-test model attributes the increase of $\overline{RT}$ with $\bar{n}_{tests}$ to the duration of an increasing number of matches. The rates of increase with $\bar{n}_{tests}$ of $\overline{RT}_{diff}$ and $\overline{RT}_{same}$ must therefore be equal to each other and to β ms per test, where β is the mean duration of a matching test. We saw this in equations 9.13 and 9.14 and in the parallel fitted lines of figure 9.7C. As you can see in that figure, the rate of increase in the data for R_{same} is far less than the rate for R_{diff}: instead of a slope of 60 ms per test for R_{diff} (which provides the estimate $\hat{\beta} = 60$ ms), the best-fitting linear function for R_{same} (also shown in the figure) has a slope of 18 ms per test. Another way to think about this is to consider the effect on $\overline{RT}_{same}$ of n_{rel} (which we manipulate) rather than n_{tests} (which we predict), as shown in figure 9.7B. Rather than growing twice as fast with n_{rel} as $\overline{RT}_{diff}$ does for $n_{diff} = 1$, the $\overline{RT}_{same}$ data grow more slowly (and perhaps even nonlinearly). Instead of $slope_{same} : slope_{diff} = 2:1$, the slope ratio of linear functions fitted to the data is almost 1:2. Thus, whereas the complex structure of the R_{diff} data are nicely consistent with the sequential mechanism (even when strong constraints are added), the relationship between $\overline{RT}_{diff}$ and $\overline{RT}_{same}$ is not at all consistent with it. This inconsistency has played a major role in attempts to explain how objects are compared for same-different judgments.

Similar difficulties for sequential testing are found in data for geometric patterns. As shown in the data in figure 9.3A for $n_{rel} = 3$ from Hawkins (1969), we have $\overline{RT}_{diff} = 524$, 430, and 386 ms for $n_{diff} = 1$, 2, and 3 (predicted $\bar{n}_{tests} = 2$, 1.33, and 1), respectively. These values give an estimate of the duration of a single test of $\hat{\beta} = 524 - 386 = 138$ ms. Because, for R_{same}, $n_{tests} = n_{rel}$, we expect $\overline{RT}_{same}$ to increase by 2×138 ms $= 276$ ms as n_{rel} is increased from $n_{rel} = 1$ to $n_{rel} = 3$. Instead, Hawkins (1969, experiment 1, shown in table 9.6) found $\overline{RT}_{same} = 443$, 465, and 481 ms for $n_{rel} = 1$, 2, and 3, respectively, a range of only 38 instead of 276 ms. Thus, although $\overline{RT}_{same}$ does increase with n_{rel}, the rate of increase is far too small, relative to the rate at which $\overline{RT}_{diff}$ increases.

3. *Linearity of $\overline{RT}_{same}$ versus number of relevant features.* The linearity that we expect of the function relating $\overline{RT}_{diff}$ to $\bar{n}_{tests}$ is beautifully borne out by the data in figure 9.7C, but the data for R_{same} (for which $n_{tests} = n_{rel}$) are concave up. Though the degree of concavity seems small, it is

reliable, and has been found in another variant of the experiment (Bamber 1972).

4. *The effect on $\overline{RT}_{\text{same}}$ of adding relevant matching attributes that are more discriminable.* Even if we ignore their relation to the R_{diff} data, the structure of the $\overline{RT}_{\text{same}}$ data presents problems for the sequential mechanism. Let F and S denote features for which matching tests are fast and slow, respectively. Consider the relation between $\overline{RT}_{\text{same}}(f,s)$ (the mean RT when F and S are both relevant and both match) and $\overline{RT}_{\text{same}}(s)$ (the mean RT when only the slower-tested feature is relevant and it matches). (In a similar argument above, associated with equations 9.21 and 9.22, we considered the effect on R_{diff} of adding relevant attributes that are *less* discriminable; here we consider the effect on R_{same} of adding relevant attributes that are *more* discriminable.) When both features are relevant, they must all be tested; thus for sequential testing, we expect

$$\overline{RT}_{\text{same}}(f,s) > \overline{RT}_{\text{same}}(s). \tag{9.23}$$

Next let us consider tests in parallel. For parallel variants 1 and 2, we expect $\overline{RT}_{\text{same}}(f,s) = \overline{RT}_{\text{same}}(s)$. Because the overlap in variants 3 and 4 implies that on some occasions $\beta_S < \beta_F$, we expect $\overline{RT}_{\text{same}}(f,s) > \overline{RT}_{\text{same}}(s)$. For parallel tests in general, we therefore expect

$$\overline{RT}_{\text{same}}(f,s) \geq \overline{RT}_{\text{same}}(s). \tag{9.24}$$

Both theories thus lead to a prediction that seems plausible: requiring additional tests, even if these are faster, will either slow the response or leave it unaffected. Surprisingly, where this comparison has been made for experiments using geometric patterns, the opposite result was obtained: $\overline{RT}_{\text{same}}$ is *shortened* when an easy attribute is added to attributes that are relevant and that must be tested. Some pertinent data from Hawkins 1969, also discussed in Nickerson 1972, are shown in table 9.6. The table shows that $\overline{RT}_{\text{same}}(h,f,c) < \overline{RT}_{\text{same}}(h)$ and, similarly, $\overline{RT}_{\text{same}}(f,c) < \overline{RT}_{\text{same}}(f)$. (You will discover two additional similar inequalities in the table that also violate equation 9.24.) This finding, *added-attribute facilitation* of R_{same}, is inconsistent with both sequential and parallel theories. Despite these serious violations of both theories, if we consider only the overall means (last column of table 9.6), we find that $\overline{RT}_{\text{same}}$ increases with n_{rel} as in Bamber's data, which is consistent with both theories; and these overall means present problems for the sequential-test theory only when the R_{diff} data are considered together with them (as discussed above). A coarse analysis can obscure important effects and thus support a theory, while a finer-grained analysis of the same data may not.

The interpretation of added-attribute facilitation is controversial. Hawkins (1969) argued that it results from subjects not completing all of the n_{rel} tests, even when they should have. (This might be described as an

Table 9.6
$\overline{RT}_{same}$ data (in ms) from Hawkins 1969, experiment 1, for various combinations of relevant attributes: *H* (height or size), *F* (form or shape), and *C* (color)

n_{rel}	Relevant feature(s)	$\overline{RT}_{same}$	Mean $\overline{RT}_{same}$
1	H	517	450
	F	455	
	C	377	
2	H,F	505	465
	H,C	468	
	F,C	421	
3	H,F,C	481	481

artifact in the experiment; see appendix 1 on trading accuracy for speed.) Such incompleteness would, of course, produce errors on some trials, and Hawkins claimed that the error patterns support his explanation. However, Nickerson (1972, 307) disagreed with Hawkins's interpretation, based on details of the relation between $\overline{RT}$ and error rate. Note (figure 9.3) that for the conditions in Hawkins's experiment for which the error rates are known, they were unusually high. This debate, still unresolved, exemplifies the difficulties in interpreting *RT* data in the presence of high error rates, with models that assume error-free performance and that are not designed to explain error rates along with *RT*s. It may be important that Hawkins selected attributes that differed widely in discriminability, which might tempt subjects to perform incomplete analyses. Because of the controversy, I ignore this phenomenon in what follows.

9.3.2 Parallel Tests Revisited

In the data considered thus far, we have seen that, for R_{diff}, processes with independent parallel tests, even after being elaborated, are not as successful as processes with sequential tests, which work well. Now we have seen that those same sequential mechanisms fail for R_{same}. The strongest argument against them is that, even though $\overline{RT}_{same}$ grows with n_{rel} as required, the observed rate of growth is too low relative to the effect of n_{rel} on $\overline{RT}_{diff}$. We have also seen that the observed growth is at an increasing rate (concave up) in Bamber's data (1969, 1972), whereas the prediction is of linear growth. Thus the sequential model fitted to $\overline{RT}_{diff}$ explains neither the size nor the shape of the effect of n_{rel} on $\overline{RT}_{same}$. The failure of sequential tests for R_{same} presents a serious obstacle to providing a coherent account of behavior in the object-comparison experiment.

In considering how to cope with this conflict it would help to know how the R_{same} responses could be handled if **FT** (figure 9.6) were a parallel mechanism.

In section 9.2.6, we saw that such a mechanism, alone, cannot explain the linear effect of n_{rel} on $\overline{RT}_{diff}$; to do so, we had to assume an effect of n_{rel} on a processing stage other than **FT**. I suggested the hypothesized encoding process, **E** (figure 9.6), as a plausible locus, which gave us the augmented parallel-test model. Because the required response is not known until the **FT** process occurs, it is reasonable to believe that any effect of n_{rel} on $\overline{T}_e$ is common to the two responses. The effect of n_{rel} on RT_{same} must then be the sum of its response-independent (common) effect on $\overline{T}_e$, together with its effect on $\overline{T}_{ft}$(same). The effect of n_{rel} on $\overline{T}_e$ is expressed by the first term ($\hat{b}n_{rel}$) on the right-hand side of equation 9.18; for Bamber's data, we found $\hat{b} = 27.0$ ms. This would also be the effect of n_{rel} on $\overline{RT}_{same}$ if its effect on $\overline{T}_{ft}$(same) was nil. The broken line in figure 9.8C is a linear function with slope 27 ms fitted (by least squares) to Bamber's $\overline{RT}_{same}$ data (1969).

In considering parallel mechanisms for R_{diff} in section 9.2.7, we looked at four variants that differ in the variability of the mismatching test durations and in their equality across features (or elements). Recall that in such a mechanism, whereas T_{ft}(diff) is the duration of the fastest mismatching test, T_{ft}(same) is the duration of the slowest matching test. The variants therefore apply to mismatching tests for R_{diff}, but to matching tests for R_{same}.

Let us consider the size and shape of the effect of n_{rel} on $\overline{RT}_{same}$. For all variants (presented below in a different order from that in section 9.2.7), we shall see that $\overline{T}_{ft}$(same) is either constant or increasing with n_{rel}. The size of the effect of n_{rel} on $\overline{RT}_{same}$ is therefore at least as great as the size of its effect on $\overline{T}_e$, shown by the broken line. The shape of the effect of n_{rel} on $\overline{RT}_{same}$ will be in between the shape of its effect on $\overline{T}_{ft}$(same) and the linear shape of its effect on $\overline{T}_e$. This means that downward (or upward) concavity of $\overline{T}_{ft}$ will produce corresponding downward (or upward) concavity of $\overline{RT}_{same}$.

9.3.2.1 Parallel Variant 1: Equal Fixed Test-Durations

This variant (figure 9.9A), which produces a nil effect of n_{rel} on $\overline{T}_{ft}$(same), is implausible because the duration of a matching test is likely to fluctuate from trial to trial, as for mismatching tests (comment 13).

9.3.2.2 Parallel Variant 4: Variable Test-Durations with Equal Means and Identical Distributions

This variant (figure 9.9D) adds plausibility. As mentioned earlier, while equality of the means across attributes is unlikely to apply to geometric

patterns without adjusting them carefully, equality of the means across locations may be a reasonable approximation for letter-string patterns. Assume, as we have earlier, that the duration of a matching test for a given feature or element not only varies from one such test to another (i.e., from trial to trial), but also that such variation for one feature is independent of the variation for another. As we did for R_{diff}, we can think of the tests on a trial as being runners in a competition, but here, because all the matching features must be tested, it is the slowest loser, not the winner, whose running time is analogous to $T_{\text{ft}}(\text{same})$. Just as there is statistical facilitation for the winner of a race (the more randomly drawn runners there are, the shorter the winner's time, on average; section 9.2.5.1), so there is "statistical inhibition" for the slowest loser (the more runners, the longer the slowest loser's time, on average). As in the case of facilitation, the size of the statistical-inhibition effect depends on the variability of the relevant test durations and not on their means. This contrasts with the effect of n_{rel} on $\bar{T}_{\text{ft}}(\text{same})$ for sequential testing, which depends on the mean test duration and not on its variability.

> *Comment 16: Demonstrating statistical inhibition.* This can be done with dice. The value, V, that comes up when a die is rolled (a random choice among the values 1, 2, ..., 6) can be regarded as the duration of a matching test. (Or this value can be regarded as the variable part of the duration; for example, the set of equiprobable durations might be $20 + V$, namely, 21, 22, 23, 24, 25, and 26 time units.) First, roll one die several times, note down the values you get, and average them. (With enough rolls, the mean should approach 3.5.) Next, roll two dice at a time, note down the maximum for each roll, and average the set of maxima. Continue this with three and four dice. Simulating on a computer the equivalent of rolling one to four dice 1,000 times, I obtained the following values for the average maximum "durations": $d_1 = 3.54$, $d_2 = 4.66$, $d_3 = 5.22$, and $d_4 = 5.62$. Clearly, the maximum grows: statistical inhibition. Note also that it grows at a diminishing rate. For example, $d_3 - d_2 = 0.56$ is only half of $d_2 - d_1 = 1.12$. The same dice experiment can be used to illustrate statistical facilitation, which, because of the symmetry of the "duration" distribution associated with a single die, is symmetric with inhibition. For the average minimum "durations," I obtained $d_1 = 3.51$, $d_2 = 2.33$, $d_3 = 1.79$, and $d_4 = 1.42$.

For independent parallel tests with identically distributed durations, what can we say in general about the form of the increase of $\bar{T}_{\text{ft}}(\text{same})$ with n_{rel}? Bamber (1972, appendix) has provided an ingenious proof that, for all the distribution shapes we might seriously consider, the function that relates the average maximum value of a (randomly sampled) set of

durations to the size of that set must be concave down. That is, it decelerates, or demonstrates "diminishing returns" of increasing the number of runners in the race, as in the dice example of comment 16. In contrast, the *upward* concavity shown by the $\overline{RT}_{\text{same}}$ data in figure 9.7C is reliable (tends to be shown by the data from most or all subjects), even when compared to a linear increase (a conservative test, because of the model's prediction of downward concavity). And similar experiments have produced similar results (Bamber 1972). The observed shape of the effect of n_{rel} on $\overline{RT}_{\text{same}}$ thus provides evidence against independent parallel tests with identical distributions for R_{same}.

Just as we found in section 9.2.7.4 for the shape of the effect of n_{diff} on $\overline{RT}_{\text{diff}}$, my computer simulations show that the degree of downward concavity (the shape of the effect of n_{rel} on $\overline{RT}_{\text{same}}$) depends on the shape of the distribution of test durations. (For example, if the duration β is long most of the time, and occasionally short, that is, negatively skewed, the concavity is greater than if β is short most of the time, and occasionally long, that is, positively skewed.) Given that the observed effect of n_{rel} on $\overline{RT}_{\text{same}}$, and hence (for the augmented model) on $\overline{T}_{\text{ft}}(\text{same})$, is concave up rather than down, one approach to fitting the model is to determine the smallest amount of downward concavity, as we vary the distribution over a set of plausible possibilities. It was the exponential distribution of β that produced the least concavity in my simulations, the same distribution that we encountered for mismatches in section 9.2.7.4.[35] For this distribution, the relation between $\overline{T}_{\text{ft}}(\text{same})$ and n_{rel}, shown in figure 9.10, departs considerably from linearity. If we let d_k denote $\overline{T}_{\text{ft}}(\text{same})$ for $n_{\text{rel}} = k$, then $(d_3 - d_2)/(d_2 - d_1) = 0.7$ and $(d_4 - d_3)/(d_2 - d_1) = 0.5$. For a linear function both of these ratios would be 1.0, of course.

Despite the difficulty of fitting the *shape* of the n_{rel} effect, it is instructive to consider the *size* of the effect. Just as for the effect of n_{diff} on $\overline{T}_{\text{ft}}(\text{diff})$ (section 9.2.7.4), the size of the n_{rel} effect on $\overline{T}_{\text{ft}}(\text{same})$ depends on the spread of the test-duration distribution. Although we know too little about how individual tests are carried out to be confident of any particular relationship between the spreads $\text{sdev}(\beta)$ and $\text{sdev}(\gamma)$, a plausible possibility is that they are approximately equal. Given equal spreads, our "prediction" of $\text{sdev}(\gamma) = 81$ ms (section 9.2.7.4), based on the size of the effect of n_{diff} on $\overline{RT}_{\text{diff}}$, is applicable to $\text{sdev}(\beta)$. Simulation shows the resulting increments in $\overline{T}_{\text{ft}}(\text{same})$ as n_{rel} increases from 1 to 2, 2 to 3, and 3 to 4 to be 41, 27, and 23 ms, respectively. Adding these values to the estimated effect of n_{rel} on $\overline{T}_e$, with the corresponding increments 27, 27, and 27 ms, and fitting the resulting values to the $\overline{RT}_{\text{same}}$ data by least squares, gives the solid curve in figure 9.8C. The size of the effect on $\overline{T}_e$ alone (broken line) is too large, relative to the data, and is a lower bound on the size of the effect on $\overline{RT}_{\text{same}}$ produced by the model, a bound that is

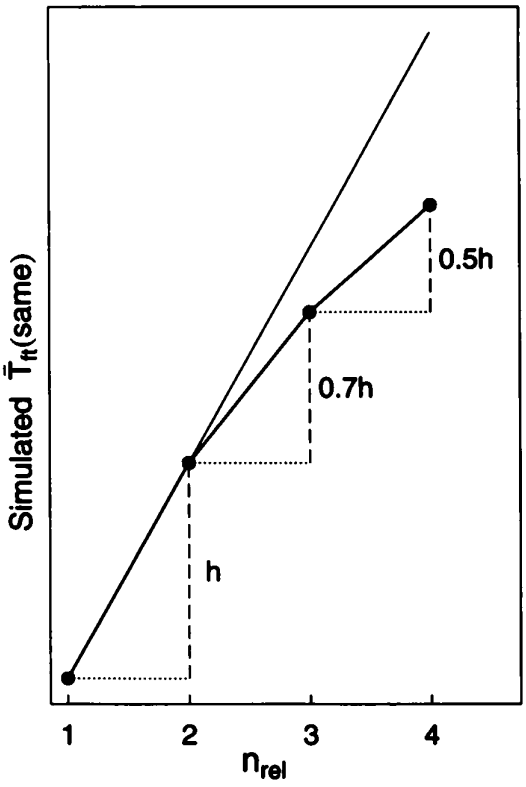

Figure 9.10
Shape of effect of n_{rel} on $\overline{T}_{ft}$ (same). This was obtained by determining the mean, $d_{n_{rel}}$, of the simulated longest of n_{rel} durations randomly sampled from an exponential distribution, and plotting it as a function of n_{rel}. In terms of section 9.2.7.4, the shape of the effect is 1.00, 0.70, 0.50. Also shown, for comparison, is a linear function (with shape 1.00, 1.00, 1.00).

achieved only by the implausible variant 1, in which n_{rel} has no effect on $\overline{T}_{ft}$. For variant 4 (as shown by the solid curve) and, as we shall see, also for variants 2 and 3, the effect predicted by the model is still greater.

9.3.2.3 Parallel Variant 2: Unequal Mean Test-Durations with Limited Variability

Given this variant (figure 9.9B) and an appropriately designed experiment, $\overline{T}_{ft}$(same) is guaranteed to grow with n_{rel}. Let the "feature ensemble" be all the features (or letter positions) ever relevant in the experiment. By "appropriately designed," I mean an experiment that is *balanced* over features in the following sense: for each value of n_{rel}, each member of the ensemble is equally likely to be one of the features. That is, for each n_{rel} value in the experiment, all possible subsets of n_{rel} features from the ensemble must be used, and must contribute equally to the $\overline{RT}$. Table 9.7 provides an example, where the feature ensemble consists of A, S, and D, and where fixed test-duration values have been used rather than variable

ones; with variability, the effect of n_{rel} will grow, due to statistical inhibition.

The example shows that $\bar{T}_{ft}$(same) is given by the duration of the slowest matching comparison for the set of relevant features. Again we see that if we sample randomly among a set of durations that are not all equal, the mean value of the maximum of the sample grows with the size of the sample. In the case of $n_{rel} = 2$, for example, $\bar{T}_{ft}$(same) is given by the mean of the maximum of a random sample of two feature-test durations. In such a sample, the pairs (β_A, β_S), (β_A, β_D), and (β_S, β_D) appear with equal likelihood; thus $\bar{T}_{ft}$(same) is given by the mean of three maxima: $\max\{\beta_A, \beta_S\}$, $\max\{\beta_A, \beta_D\}$, and $\max\{\beta_S, \beta_D\}$. (This is another instance of statistical inhibition: merely increasing the number of tests, without systematically changing the duration of any test, nonetheless slows the process.) And the growth is again at a diminishing rate, as we have seen for variant 3.

9.3.2.4 Parallel Variant 3: Variable Test-Durations with Unconstrained Means

We do not change the picture very much by going to this variant (figure 9.9C). Assuming that the durations fluctuate relatively independently, then whether $\bar{\beta}_A$, $\bar{\beta}_S$, and $\bar{\beta}_D$ differ or not, the statistical inhibition mentioned above will occur: the larger the set of features, the greater will be the maximum test duration. Furthermore, the extent of statistical inhibition will in general increase as the variability of the durations increases.

In conclusion, just as for the sequential-test model fitted to the R_{diff} data, when the parallel-test model is augmented to explain those data, it can explain neither the size nor the shape of the effect of n_{rel} on $\overline{RT}_{same}$.

9.4 Two-Process Mechanisms and Holistic Stimulus-Comparison

9.4.1 Separate Mechanisms for "Same" and "Different" Responses, and Their Temporal Arrangement

Whereas a simple process P_{diff} with self-terminating sequential tests can explain much of the $\overline{RT}_{diff}$ data, we have now seen that the same process cannot also account for the $\overline{RT}_{same}$ data. We also found that a parallel-test model, augmented to explain the $\overline{RT}_{diff}$ data, has similar difficulties explaining $\overline{RT}_{same}$. This has forced investigators to consider that two separate processes, P_{diff} and P_{same}, with different properties, underlie the two responses, despite the complexity of such a theory. The P_{same} process has been assumed either to deal with the stimulus "as a whole" or to deal with it analytically, but by using parallel rather than sequential tests.

Why might two different processes be used? One possibility is that

Table 9.7
Example of the effect of n_{rel} on $\overline{RT}_{same}$ in a parallel-testing process with unequal feature-test durations (in ms)

n_{rel}	Relevant feature(s)	Test durations	Longest duration	Mean longest
1	A	$\beta_A = 100$	100	200
	S	$\beta_S = 200$	200	
	D	$\beta_D = 300$	300	
2	A,S	100, 200	200	267
	A,D	100, 300	300	
	S,D	200, 300	300	
3	A,S,D	100, 200, 300	300	300

there is some process that is especially efficient at detecting the relation between two identical stimuli. But why would an efficient process that can make one of the decisions (R_{same}) not be used to make the other one by default—on trials on which P_{same} does not generate a "same" decision—rather than leaving the R_{diff} to a slower mechanism?

> *Comment 17: Assumption of optimality.* By asking this question, we reveal another implicit assumption, or at least a starting principle, that lies behind much of our thinking. Psychologists tend to think that because of learning, evolution, or both, people tend to use mental strategies that are efficient in some sense, and perhaps optimal, in relation to an assumed set of mental resources. If a theory claims that people do otherwise, then the theory is suspect, and needs especially strong support.

Bamber (1969) answered this question by pointing out that making R_{diff} by default would require waiting beyond the time it would normally take on "same" trials for P_{same} to produce its "same" decision. If this time was variable, then the wait would have to extend beyond the *slowest* such decision. Waiting this long might actually be less efficient than using P_{diff}. Another possible answer, proposed by Kreuger (1978), is that P_{same} is, indeed, fast, but is also prone to error; in particular, given matching stimuli, it sometimes fails to detect the match. On such trials, if R_{diff} were made by default—that is, because P_{same} failed to detect sameness—it would be made in error. To avoid such errors, the execution of R_{diff} is made to require the completion of a slower and more accurate P_{diff} process, which would presumably also generate those of the R_{same} responses not initiated by P_{same}.[36]

According to one such two-process account, the process P_{diff} that generates R_{diff} *follows* the process P_{same} that, on (most) R_{same} trials, generates

the response and ends the trial. On R_{diff} trials, P_{same} would go to completion without generating a response before P_{diff} began (which might partially or wholly explain the brevity of $\overline{RT}_{same}$ relative to $\overline{RT}_{diff}$). According to another two-process account, P_{diff} and P_{same} operate in parallel, with just one of them initiating a response, as soon as it is completed.

How might these sequential and parallel arrangements of the hypothesized P_{same} and P_{diff} be distinguished experimentally? Suppose the arrangement of P_{same} and P_{diff} were sequential, with P_{same} first. If we found a variation in the conditions of the experiment (a *factor*) that increased $\overline{RT}_{same}$, and we could plausibly argue that it did so by increasing the duration of P_{same} on R_{same} trials, then raising the level of that factor might also increase $\overline{RT}_{diff}$, and by the same amount.[37] On the other hand, if P_{same} and P_{diff} occurred in parallel, then we might be able to find one or more factors that change RT_{same} but not RT_{diff}, and vice versa. Egeth and Blecker (1971) discovered that familiarity of orientation of the patterns to be compared (which were letters presented singly or in strings of three) influences $\overline{RT}_{same}$, but not $\overline{RT}_{diff}$, providing some support for a two-process theory in general, and, in particular, for the version in which P_{same} and P_{diff} operate in parallel.

9.4.2 The Nature of the Sameness-Detection Process

Among the many issues that remain about how we compare objects, the nature of the sameness-detection process is probably the one about which least is known. If P_{same} differs from P_{diff}, what sort of process might it be? Could it be a parallel feature-testing process? Could it be a process with sequential tests, but one where the tests were carried out at a faster rate than in P_{diff}? The form of the function that relates $\overline{RT}_{same}$ to n_{rel} (concave upward, or accelerating) and also the possibility of added-attribute facilitation (table 9.6) argue against both of the simple sequential and parallel feature-test theories that we have been considering, whether or not the parallel model is augmented by an effect of n_{rel} on $\overline{T}_e$. Together with the "fast same" phenomenon, these difficulties have led researchers to consider the possibility that rather than being analytic—that is, based on decomposing the display into features (or elements)—P_{same} is holistic—that is, based on comparison of the two patterns as wholes, or gestalts.[38] As with many theoretical ideas, this one requires better definition, or elaboration, to make it predictive enough to be testable.

One elaboration that has been considered is to add the assumption that holistic comparison can be used to produce R_{same} only if the two stimuli are identical. Indeed, Bamber (1969) referred to P_{same} as an "identity reporter," and suggested (Bamber 1972) that it compares visual images of the two stimuli. Given this possibility, Bamber reasoned that it should be

possible to "turn off" an identity reporter by using different fonts for the letters in the two letter strings to be compared, and by requiring the judgment to be based on nominal identity rather than physical identity. Although the new results differed in some respects from those in Bamber's earlier experiment (1969), they shared with the early results those properties that violated the one-process theory of sequential self-terminating tests. Similarly, Miller and Bauer (1981) reasoned that an identity reporter that operated on relatively unprocessed stimulus representations could not be used to make accurate "same" decisions when the stimuli to be compared differed on attributes that had been defined as irrelevant (see section 9.1.1 for an example). Yet under conditions of successive presentation of two geometric stimuli, they found very little effect of the presence of irrelevant differences, or their number. Thus it appears that P_{same} cannot be an identity reporter operating on relatively unprocessed stimulus images.

Another approach to making the idea of "holistic" comparison more precise, so that it can be tested, was suggested by Smith and Nielsen (1970), who proposed that because (1) only one such comparison would be made, regardless of the value of n_{rel}, it follows that (2) there would be no effect of n_{rel} on $\overline{RT}_{same}$. One of the conditions in their experiment on comparison of faces produced results consistent with this claim,[39] but it is supported by the data of neither Hawkins (1969) for geometric forms nor Bamber (1969) for letter strings (see table 9.6 and figure 9.7C, respectively). On the other hand, even if proposal 1 is reasonable, proposal 2 does not necessarily follow from it: we know that $\overline{RT}_{same}$ for decisions based on single attributes (and hence, presumably, single comparisons) varies with the discriminability of the values of those attributes, as shown, for example, in table 9.6, and it seems quite possible, by analogy, that discriminability of stimuli apprehended holistically would be influenced by n_{rel}.

In another attempt to add clarity to these issues by sharpening some of the concepts, Miller (1978) elaborated the idea of an "analytic comparison process." He then used a test of this more precise idea to question whether even P_{diff} was analytic. Miller suggested that, in an analytic process, match versus mismatch decisions are made separately about different attributes. It is then these separate binary decisions (rather than the strengths of the sources of evidence that enter into the decisions) that are combined across attributes to control the response. By combining this definition of "analytic" with the assumption that the response process is ballistic (section 9.2.4.6), Miller derived powerful implications for details of the RT_{diff} data, implications he found to be violated in two visual-comparison experiments with geometric patterns. Insofar as "holistic" is defined as "not analytic" (and insofar as Miller's definition and assumption are acceptable),

these findings support the idea of a holistic comparison process underlying even R_{diff}, for geometric patterns.

> *Comment 18: Information in details of the* RT *distribution.* Let RT_o be a particular value of RT. Define a "short RT" as any $RT \leq RT_o$. Miller showed that when combined with the ballistic assumption, an analytic process (in his sense) requires the proportion of short RTs when $n_{diff} = 2$ not to exceed the sum of the proportions of short RTs for the two corresponding $n_{diff} = 1$ conditions. And this must be true for any choice of RT_o. When he applied this test, Miller found that for small and medium values of RT_o there were too many short RTs for $n_{diff} = 2$. That is, adding a second feature difference speeded the response too much. It is notable that this test makes use of details of the full distributions of RT data in each of three conditions, not merely the mean RTs. Increasingly, such details of the data are being found useful for testing alternative theories.

(Whether Miller's test would be violated by data from a letter-string experiment, in which the quantitative support for an analytic process is especially convincing, remains to be seen.) One difficulty for the interpretation of Miller's test is that it depends on the assumption of a ballistic response process, and it is hard to find independent evidence supporting this assumption.

For discovering how information is integrated from several aspects of a stimulus, without being forced to assume a ballistic response process, the measurement of the speed of decisions (in RT studies) may be usefully supplemented by the measurement of their accuracy in conditions without time pressure. Stimuli must be less than perfectly discriminable for accuracy measurements to be useful, however, and we have to keep in mind the possibility that conclusions from the study of such stimuli may not be generalizable to stimuli presented clearly. In a series of important accuracy studies, Shaw (1982) distinguished between combining binary decisions across attributes ("second-order integration"), which corresponds to Miller's "analytic process," and combining continuous, graded strengths of evidence ("first-order integration"). Data that Shaw gathered from a range of experiments favor the former.

9.5 Concluding Remarks

I began this chapter by asking how we decide whether something we see is a particular object. This question can be approached in various ways; the object-comparison experiment is appealing because it appears simple, perhaps invoking relatively few mental mechanisms and dis-

couraging alternative strategies. Such simplicity would facilitate analysis of the underlying processes. One basis for the apparent simplicity is that the task seems to be essentially visual: verbal encoding of the patterns appears not to be required, and the memory load seems minimal. (It would be desirable to check these intuitions with suitable experimental tests.) A second basis is that the responses that the subject must make remain the same as we vary the complexity of the stimuli or the number of elements they contain. In an alternative approach to visual pattern perception, in which objects would be identified rather than compared, such response invariance (which may simplify analysis of the underlying processes) might not be easy to achieve.

Much of the discussion considered how two alternative theories—parallel and sequential testing—could account for the R_{diff} data, in particular, for the effects on $\overline{RT}_{diff}$ of the number of relevant features (or elements), n_{rel}, and the number of those that differ, n_{diff}. One of the impressive aspects of the sequential theory is that it explains the full effects of these two factors, as well as the way they modulate each other, by means of a single process in which they combine to determine the number of required tests. To confront the theories with reaction-time data, we first had to elaborate and sharpen the theories to a surprising extent, to make them quantitatively specific. We saw that to have any hope of dealing with the $\overline{RT}_{diff}$ data the parallel-test theory has to be seriously augmented. Even with this revision of the parallel theory, the sequential theory still has an advantage, but the advantage is relatively small. In further tests, it will be helpful to incorporate additional experimental variations explicitly directed at some of the properties that might distinguish sequential and parallel mechanisms. Also useful will be analyses of aspects of the RT data other than their means, such as how the variability of RT_{diff} depends on n_{rel} and n_{diff}.

The sequential and parallel models developed for the R_{diff} data both run into serious trouble in explaining the R_{same} data. Unappealing as it is to introduce such complexity, we are forced to conclude that the two responses are generated by different processes, P_{diff} (about which we know a good deal) and P_{same} (about which much more needs to be learned).

As you have read this chapter, you will have thought about some of the important issues that arise in working with reaction-time data, developed intuitions about serial and parallel processing mechanisms, learned how theories of such mechanisms must be elaborated to make them testable, and, in general, have had some practice in making inferences from behavioral data to underlying mental mechanisms. Along the way, I have shown you how questions about the object-comparison experiment have been sharpened, described a few of the methods developed to answer them, and summarized some answers. Starting with Egeth 1966 and Bamber

1969 there has been much progress, but intriguing puzzles remain to be solved.

Appendix 1: Error Rates and the Interpretation of Reaction-Time Data

The kinds of inference discussed in this chapter depend on quantitatively specified effects of experimental factors on *RT*, factors such as n_{rel} and n_{diff} in Bamber's experiment (1969). On what basis can we take seriously the quantitative details? We know that other variables that are not the focus of these experiments can influence *RT*—variables such as the amount of practice, time of day, and the subject's level of motivation. It is critical that either such variables change minimally in the course of the experiment, or that they change in a way that is not correlated with ("confounded" with) the factors that interest us.[40] Indeed, experimental design is largely concerned with such issues.

It is known that under enough time pressure, subjects can be induced to trade accuracy for speed, and that they can adopt different trading relations under different conditions. Might that be happening in Bamber's experiment (1969)? A first glance at figure 9.4B suggests that we may be in trouble: the error rate changes systematically with both n_{rel} and n_{diff}. If this variation in error rate is a reflection of the trading of accuracy for speed, then it is possible that the *RT* pattern that we see is a distorted one, influenced by a trade-off strategy that, for example, varies with n_{rel}. (Because the two letter strings are presented successively, separated by a comfortable time interval, information about n_{rel} is available before the *RT* clock starts, unlike n_{diff}, which could permit subjects to adjust their "strategy" for performing the task in response to n_{rel}.)

One simple way in which subjects might trade accuracy for speed would be to guess randomly on some proportion of the trials, rather than taking the time needed to process the stimulus (beyond merely detecting that it has occurred). In a typical object-comparison experiment, a random half of such "fast guesses" would be correct, which creates a second problem associated with the error rate. Even if the $\overline{RT}$s reported for a set of conditions were based on just the correct responses (as is usually done), they would include 50 percent of the fast guesses that happened to be correct; that is, they would be contaminated by data arising from a different process from the one we wish to study. If this simple trading mechanism were the only one, then we might be able to use the *RT*s on error trials to correct the contamination effect, and so estimate for each "condition" (e.g., value of n_{rel}) the proportion of guesses and the "true *RT*." On the other hand, errors might arise from more complex trading mechanisms—for example, the partial but incomplete stimulus analysis produced by ignoring one of the relevant attributes in a multiattribute experiment.

A straightforward interpretation of *RT* data is therefore challenged by two issues related to the error rate. First, the amount of contamination increases as the error rate increases. And second, variations in the error rate across conditions may indicate trading of accuracy for speed to a different extent, or according to different rules, in different conditions. For these reasons, in any experiment it is important to consider error rates along with *RT*s; this is partly why I included figures 9.3B and 9.4B. These figures also show that error rate can vary markedly across different conditions (i.e., different values of n_{rel} and n_{diff}) that may be mixed together randomly from trial to trial. The experimenter therefore needs not only to be aware of overall error rate, averaged over conditions or factor levels, but also of the rates for individual conditions.

In an experiment with "typical reaction-time instructions," subjects might be asked to "please respond as rapidly as possible, consistent with high accuracy." When I run experiments, I usually try to convey the relative importance of speed and accuracy more precisely, by using explicit payoffs. For each block of twenty trials, for example, subjects might get a

point for each hundredth of a second in their average RT, and ten points for each error; they would be asked to minimize the score. I try to make the penalty for errors sufficiently great, relative to the cost of time, so that under none of the conditions in the experiment does guessing pay.

Under different instructions, if the experimenter arranges for the cost of time to be large relative to the cost of errors, especially by introducing costs that increase abruptly when an RT deadline is exceeded, subjects can be induced to respond more rapidly and make more errors. One conclusion sometimes drawn from the observation that subjects under severe time pressure are capable of such flexibility is that they are always exercising it, even under more typical instructions.[41] If so, experimenters would have to decide what cost of errors relative to time they should impose on subjects. As I will try to explain below, it is not clear what error rate the experimenter should aim at in each condition of an experiment (e.g., for each value of n_{rel}).

If subjects are freely trading accuracy for speed (and doing so differently, or to different degrees, on different types of trial), this would create serious difficulties for the interpretation of RT data. Roughly speaking, if subjects are engaged in such trading, then this might produce unknown or even unknowable biases in the measured $\overline{RT}$ relative to the "true $\overline{RT}$," and might even raise questions about how to define the true $\overline{RT}$. The existence of such biases would interfere with our ability to compare RTs from different conditions or tasks.

Much discussion of the relation between speed and accuracy in RT experiments (e.g., Pachella 1974; Wickelgren 1977) incorporates the following three assumptions:

> 1. A subject's performance in each task or experimental condition lies on a nondecreasing "speed-accuracy trade-off function" relating accuracy, that is, percent correct, $P(c)$, to $\overline{RT}$. Two fictitious trade-off functions are shown in figure 9.11.
> 2. The subject can adopt a stable and arbitrary point on the trade-off function, and does so by estimating what the function is from the data accumulating over trials, and optimizing the chosen point in relation to the explicit and implicit payoffs for speed and accuracy. (Given the "fast guess" mechanism, for example, selection of a point on the trade-off function would be accomplished by choosing the percentage of trials on which to guess.)
> 3. If the trade-off functions for two tasks, A and B, are distinct, then they do not cross, and are therefore related by *dominance*. Figure 9.11 has been drawn so that A dominates B: for any $\overline{RT}$, $P(c)$ is greater for task A than for task B.

If performance places the subject at point a or a^* in task A and at point b^* (slower and more accurate) in task B, then this would not provide evidence of different trade-off functions; because a nondecreasing function could be found that would pass through the two points, some would say that the difference between tasks might be "due to a speed-accuracy trade-off" (on a single trade-off function). However, if the performance were represented by point a or a^* in task A and point b (slower and no more accurate) in task B, this is enough to tell us that performances in the two tasks lie on two different trade-off functions, and that the dominance relation favors A. Thus, given that the data meet such a *speed-accuracy correlation requirement* (task B is performed both more slowly and less accurately than task A) the idea that the two tasks share the same trade-off function can be rejected: some authors would claim that such a difference in mean RT is "not due to a speed-accuracy trade-off." For some purposes, especially in relation to practical questions (where the level of performance of whole tasks may be of primary interest, rather than the understanding of underlying processes) it is useful to be able to say that task B is "harder" than task A in this sense. Figures 9.3 and 9.4 show that the speed-accuracy correlation requirement is met, in general, in the experiments we have been considering. That is, conditions with slower performance also have higher error rates. Thus increasing n_{diff} ($n_{diff} \geq 1$) makes the task "easier," and increasing n_{rel}

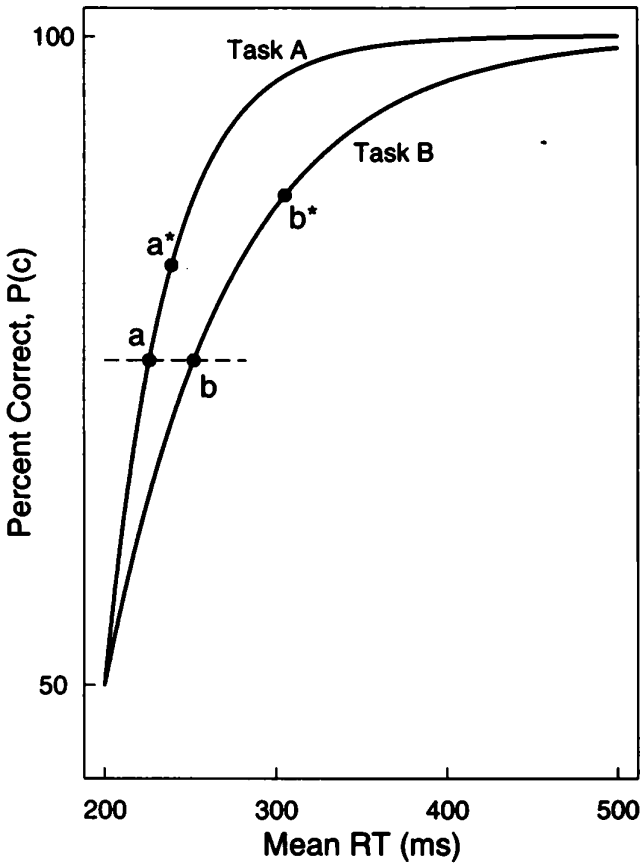

Figure 9.11
Schematic speed-accuracy trade-off functions for two tasks, A and B. Percent correct, $P(c)$, is plotted as a function of $\overline{RT}$ for each task. The function for task A dominates the function for task B. The direction of dominance can be determined by measuring pairs of points such as a (or a^*) and b on the two functions. Points a and b exemplify conditions with equalized error rates. Points a and b^* exemplify differences in $P(c)$ and $\overline{RT}$ that could be due to a trade-off rather than to more basic task differences.

makes the task "harder"; given assumptions 1 and 3, these effects cannot be "due to a trade-off," in this sense.

Because trade-off functions tend to rise more slowly at higher levels of accuracy, a large difference in $\overline{RT}$ corresponds to a small difference in error rate. This viewpoint has suggested to some that instead of collecting data under conditions of high accuracy (for which we often strive, to reduce contamination by "guesses" that happen to be correct), data should be collected under conditions of lower accuracy (where the trading function is steeper); under these conditions, it is believed that the estimated error rate would be more sensitive to position along the trade-off function, which would therefore be easier for both the subject and the experimenter to monitor.

Given the three assumptions above, it is difficult to justify *quantitative* comparisons between mean *RT*s from different conditions in an experiment: we might think that by measuring (or estimating) performance in the two tasks when error rates are *equalized* (a and b in figure 9.11) we could measure a "true" *RT* difference, but this depends on assuming that the change in task has no *inherent* effect on error rate—an assumption hard to justify. For example, suppose the change from task A (e.g., $n_{rel} = 2$) to B (e.g., $n_{rel} = 3$) is associated with an

increase in the number of operations (e.g., tests) between stimulus and response, and that each operation has some failure probability. Then even if we arranged for the failure probability *per operation* to be the same for tasks A and B, they would have different rates of *overt* (measured) error. For example, each letter test in Bamber's experiment (1969) might have some failure probability, and as n_{rel} increases for fixed n_{diff}, the average number of required tests increases. This would increase the percent of false same responses as n_{rel} increases, an effect that Bamber observed (figure 9.4B). Thus merely observing that error rate varies systematically with n_{rel} and n_{diff} should not necessarily be disturbing, despite a possible initial impression to the contrary on seeing figure 9.4B. Conversely, *equalizing* overt error rates between tasks A and B would render failure probabilities per operation *unequal*, and might not be what we want for interpretable $\overline{RT}$ data. A further complication is that the time between stimulus and response is probably occupied by more than one operation susceptible to speed-accuracy trading (for example, one or more feature tests, as well as the residual operations mentioned in section 9.2.3.1). If so, the subject might be free to select positions on several trade-off functions independently, there might be no unique trade-off function for a task, and the dominance assumption might be hard to justify.

An approach sometimes adopted is to accept the idea that trading is ubiquitous, and apply strong time pressure so as to measure several points on the empirical trading function (e.g., McElree and Dosher 1989; Meyer, Irwin, et al. 1988). Another approach, not often attempted, is to devise theories that describe both *RT*s and errors (e.g., Ashby and Maddox 1994; Ratcliff and Murdock 1976).

If we accept the ubiquity of trading, it is puzzling that we ever observe highly orderly behavior of the mean *RT* in response to task changes, as exemplified in figure 9.7B. How can such occurrences be reconciled with the idea that a range of points on the speed-accuracy trade-off function are readily accessible to the subject? Or with the idea that the subject must learn what the trading function is for a condition to make the relevant adjustments, even though the sample data usually available to the subject, especially about error rate, are very sparse. (An argument that it is the "demand characteristics"—the subject's understanding of what the experimenter desires of the data—that induce order in our data seems extremely implausible for various reasons, especially because subjects often appear unaware of the data patterns they produce.) Also puzzling is the observation of the simple and plausible orderliness of quantitative relations among error rates across conditions in some *RT* experiments, as demonstrated by Schweickert (1985).

An alternative possibility that has been given inadequate attention is that, although trading can be made to occur, it does not spontaneously occur with typical *RT* instructions of the sort described above. That is, there may be conditions under which subjects respond when they have "completed" what has to be done, in the sense of acquiring a sufficient amount of response-relevant information. (Subjects may have to produce occasional errors while learning what is sufficient.) Some support for the idea that *RT* experiments do not normally elicit trading behavior derives from results obtained by Pachella (1974), who studied subjects first under "typical" instructions, and then under instructions that induced them to trade time for errors. When the subjects returned to the "typical" instructions, their behavior was qualitatively different from what it had been originally. It seems hard to reconcile this outcome with the idea that "typical" instructions merely induce subjects to adopt one of many possible arbitrary points on the trading function. Pachella (1974, 69) suggests that "speed stress leads to a basic change in processing strategy." And Wickelgren (1977, 81) has commented, more generally, "To be completely fair, we do not know for certain that subjects' capability to achieve any degree of speed-accuracy tradeoff ... under speed-accuracy tradeoff instructions implies that this capacity is used under reaction time instructions."

I believe that ultimately we shall have to go further in developing theories for *RT* experiments that deal with errors as well as time. For the present, whatever our measure, we are

operating on a good deal of faith. The test of that faith lies principally in the orderliness, replicability, and interpretability of the data.

In my own experiments I respond to these difficulties in several ways. First, I try to devise incentive and feedback arrangements that seem not to place differential pressure on subjects under different conditions. (One approach is to randomize conditions over trials and to "charge" a subject an amount that increases linearly with $\overline{RT}$. Another approach, if different conditions are studied separately, is to reward the subject by an amount that depends on improvement relative to previous performance in that condition. I avoid incentives that depend on whether a general RT deadline is met.) Second, when I investigate a new condition I try to incorporate a replication of one or more conditions examined earlier, to make possible a consistency check. Third, I deliberately incorporate extra variables of secondary interest in experiments—variables I hope will leave the effect of interest invariant—to permit tests of generality. Fourth, I am attracted by methods that produce orderly data because I assume that nature is likely to be more regular than the effects of experimental errors, or "artifacts." On the other hand, I recognize that given the possibility of speed-accuracy trading, some of the fine detail in RT data may be spurious.

Appendix 2: Donders' Subtraction Method and Modern Variants

Suppose you wanted to know the duration of a single matching letter test, that is, the value of β in equations 9.14 and 9.15.

Experiment A: An attempt to measure β directly. One possibility would be to present letter strings that contain just a single letter ($n_{rel} = 1$) and to ask a subject to press a button if the test letter is the same as the target and not to respond otherwise (a "go/no-go" procedure). We would measure the $\overline{RT}$ on "same" trials, $\overline{RT}_{same}(1)$. Let us call this "condition 1." Because "same" and "different" trials would occur with equal frequencies, the subject would press the button on half of the trials. Before reading further, consider how well β would be estimated by $\hat{\beta}_1 = RT_{same}(1)$, and why.

By using a go/no-go procedure (respond if "same," do not respond if "different") rather than a choice procedure (one response if "same," another response if "different"), we have permitted the subject to prepare the response in advance and have probably simplified and shortened the residual operations by doing so. Nonetheless, because the RT in condition 1 would very likely still contain the durations of residual operations (α_{same}) as well as the desired test duration (β), $\hat{\beta}_1$ would not answer our question. That is, $\hat{\beta}_1$ would measure $\alpha_{same} + \beta$ (as in equation 9.7) rather than β alone. To use $\hat{\beta}_1$ to measure β, we would also need a separate estimate of α_{same} to subtract from it.

Experiment B: An attempt to measure β and α_{same} directly. A second possibility might be to modify experiment A by adding a condition 0, whose aim would be to provide a measure of just the durations of the residual operations, α_{same}, excluding the test duration. On half of the trials in condition 0, no test letter would be presented, and subjects would be asked to withhold their response. On the other half of the trials, a test letter would be presented, and subjects would have to press the button regardless of whether the test letter was the same as the target letter or different from it. In other words, subjects would be asked to respond to any letter. Before responding, subjects would have to detect the presence of the test letter, but not test its relationship to the target letter. Let us call the resulting measurement $\hat{\alpha}_0 = \overline{RT}_{detect}(0)$. The stimulus on response trials in condition 0 would be the same as in condition 1, as would be the response. And the frequencies of response and nonresponse trials in the two conditions would be the same. Before reading further, consider how well α_{same} would be estimated by $\hat{\alpha}_0$, and why.

Suppose we can assume that the $\overline{RT}_{detect}(0)$ from condition 0 includes the durations of all the operations in condition 1 except the sameness test, that is, that $\hat{\alpha}_0$ is a good estimator of

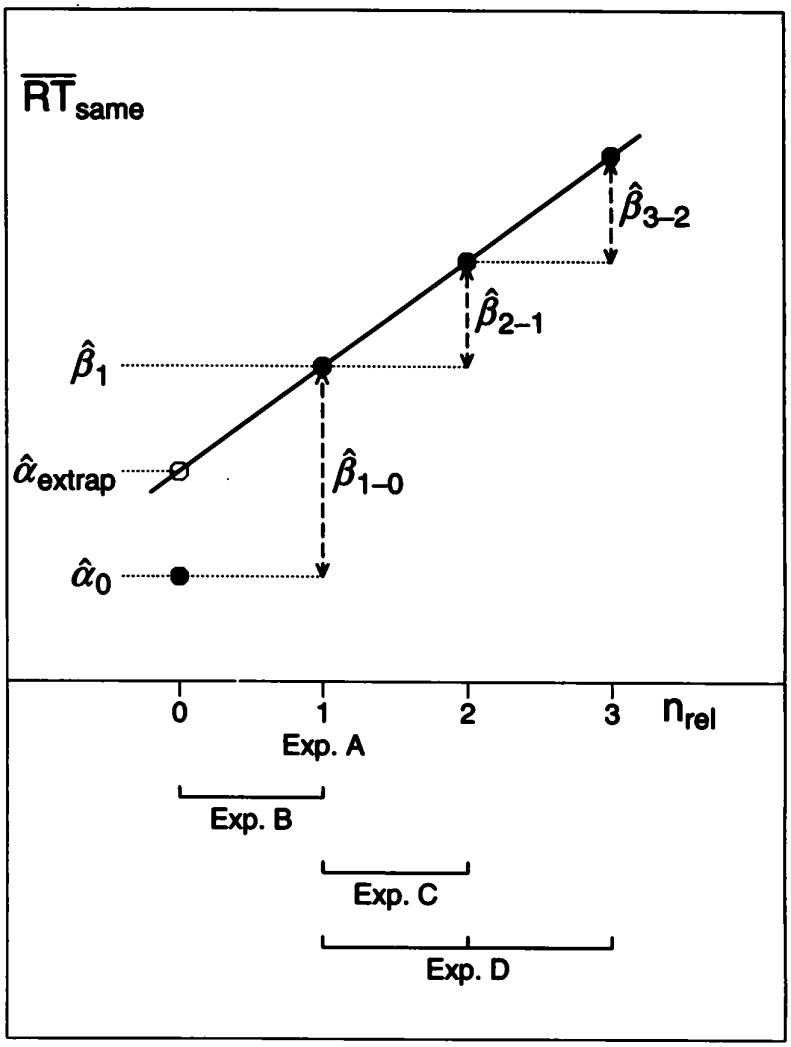

Figure 9.12
Idealized results from four alternative experiments, A, B, C, and D, designed to provide estimates of β. Conditions (tasks) are numbered 0, 1, 2, and 3, corresponding to n_{rel}. An experiment consists of one or more conditions whose results are combined in making inferences. The conditions included, by experiment, are A: 1; B: 0, 1; C: 1, 2; and D: 1, 2, 3.

α_{same}. This is an instance of an assumption of *pure insertion*, discussed in comment 3 and section 9.2.3.1. According to this assumption, changing from condition 0 to condition 1 inserts a matching letter test without changing any of the other operations; a reasonable estimate of β would then be provided by $\hat{\beta}_{1-0} = \hat{\beta}_1 - \hat{\alpha}_0 = \overline{RT}_{same}(1) - \overline{RT}_{detect}(0)$, shown in figure 9.12.

Let us step back and consider what requirements must be met for $\hat{\beta}_{1-0}$ to be a good estimator of β. The validity of estimates such as this, based on subtraction, depends on the truth of at least three propositions:

1. *A task analysis.* By "task analysis" I mean a statement of a theory or hypothesis that specifies the mental operations used to carry out a task. Because the subtraction method is applied to a pair of tasks, two such task analyses are required. For the case

of $\hat{\beta}_{1-0}$, the task analyses must include the idea that conditions (tasks) 1 and 0 differ only by the incorporation of a matching letter test in 1 that is not incorporated in 0. In other cases, where one or both conditions involve displays of $n_{rel} > 1$ letters, the idea that the comparison process is analytic (letters tested separately) rather than holistic would be incorporated in the task analysis.

2. *A stage theory.* A stage theory is a statement about the temporal organization of subsets of the operations identified in the task analysis. It asserts that subsets of one or more operations are arranged in stages. In this case, the matching letter test (one subset) and the residual operations (another subset) used to respond in condition 1 must be arranged in stages, such that the temporal epoch occupied by the letter test does not overlap the epoch occupied by the residual operations. (When processes are arranged in stages, one process begins when its predecessor is completed.)

3. *An assumption of pure insertion.* We have to assume that durations of the residual operations are not altered as we move from condition 0 to condition 1, inserting the test operation.

Before reading further, consider how you might test the validity of these propositions.

The estimation procedure in experiment B is an instance of the *subtraction method* devised by Donders (1868) and used enthusiastically during the last quarter of the nineteenth century. Later, psychologists recognized that in any application of this method, the pure-insertion assumption needed justification. (There was less explicit concern with the support needed for the other two important propositions.) In the present case, for example, it seems likely that the operation by which the test letter is encoded—one of the possible residual operations—might differ, depending on whether the letter has merely to be detected, as in condition 0, or has to be compared to the target, as in condition 1. The information about the test letter that is required by the comparison process would probably be more extensive than the information required merely for detection, in which case the duration of the encoding operation might be greater. Thus $\hat{\alpha}_0$ might underestimate α_{same}; hence $\hat{\beta}_{1-0}$ would overestimate β. The hypothetical data shown in figure 9.12 reflect these estimation biases.

Experiment C: An attempt to estimate β without measuring α_{same}. A third possibility would be to modify experiment A by adding condition 2 to condition 1. The only difference between these conditions is that condition 2 would involve letter strings of size two instead of size one ($n_{rel} = 2$), and would provide us with $\overline{RT}_{same}(2)$. Unlike condition 0, conditions 1 and 2 would both require a comparison process. In this application of the subtraction method, the difference $\hat{\beta}_{2-1} = \overline{RT}_{same}(2) - \overline{RT}_{same}(1)$ would provide the estimate of β. Before reading further, consider how good an estimate $\hat{\beta}_{2-1}$ is, and how you might test the validity of the required pure-insertion assumption.

Experiment C is likely to have the advantage that for both conditions 1 and 2 a comparison process is needed, so that the encoding process would have to be sufficiently elaborate in both conditions to be suitably informative. But a possible disadvantage is that unlike condition 0, the stimulus in condition 2 would differ from that in condition 1 (two letters rather than one). This could mean that the encoding operations would again differ between the two conditions, but for a different reason. However, we have already seen evidence (in section 9.2.4) that the duration of the encoding process in Bamber's experiment (1969) is relatively unaffected by the number of letters, $1 \leq n_{rel} \leq 4$. For purposes of the present discussion, let us suppose this to be true.

An extrapolation estimate of α_{same}. Although our attempt to measure α_{same} directly has presumably failed, because the required pure-insertion assumption relating conditions 0 and 1 was invalid, the data from experiment C permit us to get an indirect estimate of α_{same}. Because $\hat{\beta}_{2-1}$ provides an estimate of β, and $\hat{\beta}_1$ provides an estimate of $\alpha_{same} + \beta$, the differ-

ence between them, $\hat{\beta}_1 - \hat{\beta}_{2-1} = 2\overline{RT}(1) - \overline{RT}(2)$ gives us an estimate of α_{same}. (To convince yourself, write the two $\overline{RT}$s in terms of α_{same} and β.) Another way to think of this is that conditions 1 and 2 give us measures of $\alpha_{same} + \beta$ and $\alpha_{same} + 2\beta$, respectively, permitting us to estimate α_{same} indirectly, by extrapolation, rather than by devising a special condition to provide us with a direct estimate. That is, $\hat{\alpha}_{extrap}$ is the zero intercept of a line fitted to the estimates of $\alpha_{same} + n\beta$, $n = 1, 2$.

Experiment D: An elaboration of experiment C. Adding condition 3 to conditions 1 and 2 would give us another estimator, $\hat{\beta}_{3-2} = \overline{RT}_{same}(3) - \overline{RT}_{same}(2)$, along with $\hat{\beta}_{2-1}$, and would also provide us with a check on the validity of the method, as we shall see.

We have now accumulated four different potential estimators of β: $\hat{\beta}_1$, $\hat{\beta}_{1-0}$, $\hat{\beta}_{2-1}$, and $\hat{\beta}_{3-2}$. I argued that $\hat{\beta}_1$ would overestimate β, because of inclusion of residual times, α_1, and also that $\hat{\beta}_{1-0}$ might overestimate β because the encoding time in condition 0 might be reduced relative to condition 1, that is, a failure of pure insertion. If the two remaining estimators are both good ones, then we would expect agreement between the values they provide. Before reading further, consider what such agreement would require of the $\overline{RT}$ data.

Agreement between $\hat{\beta}_{2-1}$ and $\hat{\beta}_{3-2}$ means that $\overline{RT}_{same}(3) - \overline{RT}_{same}(2) = \overline{RT}_{same}(2) - \overline{RT}_{same}(1)$, or that the three $\overline{RT}$s would fall on a linear function. This is what is implied if the three propositions required for the subtraction method are valid in experiment D. Unfortunately, however, we have seen (figure 9.7B) that $\overline{RT}_{same}(n)$ is concave up rather than being linear in n. Applied to this example, the propositions require that in each change, from condition 1 to 2, and also from 2 to 3, we add a stage of processing that carries out the matching test of a single letter. That is, R_{same} has to be generated by a sequential mechanism. But we concluded earlier in the chapter that some other mechanism is at work. The subtraction method is therefore inappropriate here.

In an alternative approach to the estimation of β, we could consider experimental conditions that gave us values of $\overline{RT}_{diff}$ rather than $\overline{RT}_{same}$. For example, we might use a go/no-go procedure in which subjects responded when the letter strings differed, and made no response otherwise. Suppose we set $n_{diff} = 0$ or 1 and varied n_{rel}. Responses would be generated on the $n_{diff} = 1$ trials. From equations 9.1 and 9.2, we know that $\bar{n}_{mtests}(\text{diff}) = (n_{rel} - 1)/2$; this equation expresses part of the required theory of task operations, which also asserts that a single mismatching test is included among the operations on any comparison trial leading to a response. Again we could construct estimators based on the subtraction of two $\overline{RT}$ values, now $\overline{RT}_{diff}$ values associated with successive values of n_{rel}. $\overline{RT}_{diff}(1)$ (for $n_{rel} = 1$) would incorporate no matching tests, a single mismatching test, and residual operations. $\overline{RT}_{diff}(2)$ would add (on average) half the duration of a matching test. One desired estimator would then be $\hat{\beta}'_{2-1} = 2[\overline{RT}_{diff}(2) - \overline{RT}_{diff}(1)]$. (The prime is used simply to distinguish estimates based on R_{diff} data from those based on R_{same} data.) Another estimator would be $\hat{\beta}'_{3-2} = 2[\overline{RT}_{diff}(3) - \overline{RT}_{diff}(2)]$. In this case, the three $\overline{RT}$ values do fall on a linear function (the n_{diff} function in figure 9.7B), the two estimators provide similar values, and we therefore have support for the propositions that validate the method. Here the zero intercept of the linear function provides us with an extrapolation estimate of $\alpha_{diff} + \gamma$.

There are three important conclusions to be drawn about the subtraction method. First, in a situation with repeated operations, where there is a possibility of arranging for a controlled number of operations of the same type (here, letter tests), we have the opportunity to assess the method's validity. These arguments are most widely applied in tasks that call for search—of visual displays or memorized lists—where it may be reasonable to believe that the experimenter can control the number of repetitions. In a more traditional application of the method, we might have used only the $\hat{\beta}_{1-0}$ estimator; we would have had to search elsewhere for a validity check and might not find one that was satisfactory. (Introspection by subjects appears to have been used for this purpose during the early years of the method.)

Second, the situation with repeated operations provides us with an indirect extrapolation method to estimate the mean duration of residual operations, using conditions with two different numbers of repeats, which substitutes for the possibly flawed direct method that Donders would probably have used. Third, one possible validity test with repeated operations is a test of linearity of the measures obtained from conditions with three different numbers of repeats.

Glossary

Listed below are the main symbols used in the text, with numbers of the sections in which they are introduced, and brief definitions.

$\alpha, \beta, \gamma, \delta, \varepsilon, \theta$	(various sections)	The Greek letters alpha, beta, gamma, delta, epsilon, and theta, often used to denote constants, or parameters, in models
RT	(9.1)	Reaction time, time from stimulus onset to response detection
$\overline{RT}$	(9.1)	Mean of a set of RTs
R_{same}, R_{diff}	(9.1)	"Same" and "different" responses
n_{rel}	(9.1.1)	Number of features (or elements) that are relevant to the same-different decision
n_{diff}	(9.1.1)	Number of features among the n_{rel} that differ between objects being compared
$\overline{RT}_{same}$, $\overline{RT}_{diff}$	(9.1.4)	Average reaction times for correct "same" and "different" responses
n_{tests}	(9.1.4)	Number of feature tests associated with a response
$\bar{n}_{tests}$	(9.1.4)	Mean number of tests associated with a response
$n_{tests}(diff)$	(9.2.1)	Number of tests associated with R_{diff}
n_{mtests}	(9.2.2.1)	Number of tests of features that match
FT	(9.2.3.1)	Feature-testing process
T_{ft}	(9.2.3.1)	Duration of **FT**
E	(9.2.3.1)	Encoding process
T_e	(9.2.3.1)	Duration of **E**
α_{same}, α_{diff}	(9.2.3.1)	Mean sum of durations of residual operations associated with R_{same} and R_{diff}
θ	(9.2.3.2)	Duration of one test, when matching and mismatching test durations are equal
$\hat{\theta}$	(9.2.4)	Value of θ estimated from data
β_A	(9.2.4.4)	Duration of a matching test for feature A
γ_A	(9.2.4.4)	Duration of a mismatching test for feature A
sdev	(9.2.7.4)	Standard deviation (square root of the variance)
P_{same}, P_{diff}	(9.3.3.1)	Separate processes that might generate R_{same} and R_{diff}

Suggestions for Further Reading

Excellent introductions to the use of RT in research on human information processing are Meyer, Osman, et al. 1988 and Pachella 1974. Luce 1986, Townsend and Ashby 1983, and

Welford 1980 are advanced treatments; the first two emphasize mathematical models. Reviews of basic *RT* phenomena can be found in Smith 1968, Keele 1986, and, for earlier work, Jastrow 1890 and Woodworth 1938, chapter 14; much of Chase 1978 and Posner and McLeod 1982 are also of interest. Schweickert (1993) provides a recent review, emphasizing theoretical ideas. If you are interested in the early history of the subject, you will also enjoy the papers by and about Donders in the proceedings of the Donders Centenary Symposium on Reaction Time edited by Koster (1969), and also some of the papers by James McKeen Cattell that have been collected in Cattell 1947. Reports of recent high points in the use of *RT* to learn about human mental processes can be found in the Attention and Performance series, whose volumes have been published approximately every two years since 1967; these also contain useful tutorial reviews.

Corcoran 1971 and Reed 1973 are excellent introductions to pattern recognition. A good starting point for learning more about attempts to understand how subjects behave in visual-comparison experiments is Nickerson's fine review (1972, 301–312). You should also see Kreuger's proposed single mechanism (1978) for R_{diff} and R_{same}, and the suggested revision of Kreuger's theory by Miller and Bauer (1981), as well as reviews, theories, and experiments by Farrell (1985, 1988), Proctor (1981), and Proctor, Rao, and Hurst (1984). In an approach to analyzing the visual-comparison experiment not mentioned in the present chapter, an object is represented as a point in a multidimensional space, and the distance between two such points reflects the discriminability of the corresponding objects; see Lockhead 1972, Nosofsky 1992, and Sergent and Takane 1987.

Townsend (1990) discusses valid and invalid methods for distinguishing between sequential and parallel mechanisms, and provides a useful guide to other such discussions. Examples of the use of properties of the *RT* distribution other than its mean for understanding mental mechanisms (including properties akin to the shortest RT_{diff} mentioned in the seventh of the questions for further thought) can be found in Vorberg 1981, Yantis, Meyer, and Smith 1991, Townsend and Ashby 1983, especially chapter 8, and in Roberts and Sternberg 1993. Arguments for and against the assumption of a ballistic response process are presented by Meijers and Eijkman (1977) and Giray and Ulrich (1993).

Questions for Further Thought

9.1 *Simultaneous versus successive displays.* Figure 9.3A shows that responses are faster when the stimuli to be compared are displayed successively rather than simultaneously. Consider at least two reasons why this might be so, and how you might test them.

9.2 *Box score.* Construct a table listing aspects of the data discussed in this chapter that are favorable and unfavorable to each of the models considered. (Different sequential-test models are generated by different combinations of constraints.)

9.3 *Effect of n_{diff} in parallel testing.* The magnitude of the statistical facilitation effect in a parallel-test mechanism (illustrated in table 9.5) is influenced by test-duration variability. To show this, simplify the situation described in the table by assuming that the three attributes have identical two-point distributions. This means that you can omit table sections for S, D, (A,D), and (S,D). For the high-variability case, let the two equiprobable test durations be 50 ms and 150 ms. You should be able to show that the facilitation effect is 37.5 ms, smaller than the effect of 62 ms in the table. For the low-variability case, keep the same mean, and let the two durations be 90 ms and 110 ms; you should be able to show that the facilitation effect drops to 7.5 ms. Note whether there are "diminishing returns" in these two cases, and compare the results to each other and to those in the table. The decline in facilitation from the first case to the second illustrates the fact that if we knew the variability of the test durations, we could say something about the effect of n_{diff} on the $\overline{RT}$ produced by a parallel mechanism.

9.4 *Proof of equation 9.2.* The proof of equation 9.2 employs combinatorial probability. (For one introduction to this subject, see Feller 1968, volume 1, chap. 2.) Here are some hints for a proof. Using the language of "targets" and "nontargets," we need to consider the length of the starting *run* of nontargets in the search path. (A "run" of nontargets is an uninterrupted sequence of nontargets followed by a target.) Let R be this length. (If the first element is a target, then we define $R = 0$.) We have to determine the probability of occurrence of a starting run of each possible length, that is, the *probability distribution* of run length. There are t targets and $s - t$ nontargets in the set of s elements. The probability that $R = r$ is the probability that the first $r + 1$ elements in the search path consist of r nontargets followed by one target. This probability is given by multiplying the number of ways of choosing r nontargets from the $s - t$ nontargets by the number of ways of choosing 1 target from t targets, and dividing the resulting product by the number of ways of choosing $r + 1$ elements from the s elements. Once we have the probability distribution of R, we use it to obtain its mean, $\bar{R}$; the mean number of tests will then be $\bar{R} + 1$.

For the special case of $n_{\text{diff}} = 1$ target, and $n_{\text{rel}} - 1 = s - 1$ nontargets, the proof that $\bar{n}_{\text{tests}}(\text{diff}) = (s + 1)/2$ (as indicated by equation 9.2) is easier. We assume that the target has a probability of $1/s$ of appearing in each of the s locations in the search path. Because the number of tests required is given by the location of the target, $k = 1, 2, \ldots, s$, the mean number of tests is the mean of $1, 2, \ldots, s$. This mean is $(1 + 2 + \cdots + s)/s$, and because $1 + 2 + \cdots + s = s(s + 1)/2$, the value of the mean is $\bar{n}_{\text{tests}} = (s + 1)/2$. In an alternative proof for this special case, we consider the mean number of nontargets that precede the target in the search path. Because the target is equally likely to be the first, second, ..., last element in the path, it is preceded, on average, by half of the $s - 1$ nontargets. The mean number of elements tested is therefore $(s - 1)/2$ nontargets plus 1 target, or $(s + 1)/2$ elements.

9.5 *Special aspects of face recognition.* Evidence from brain-damaged subjects suggests that the mechanism used for face recognition may be different from the mechanisms used to recognize other objects (Farah 1992; see also chap. 3 of volume 2, this series). This makes it interesting to compare same-different judgment of pairs of faces to judgment of letter strings or geometric patterns. Smith and Nielsen (1970) ran an experiment in which subjects made same-different judgments of schematic faces, which were presented successively, with interstimulus intervals (ISIs) of 1, 4, and 10 sec. (Lengthening the ISI should reduce the subject's ability to use visual images of the two stimuli.) At all ISIs, despite a substantial effect of n_{diff} for fixed n_{rel} (an effect that increased with ISI), the effect of n_{rel} for fixed n_{diff} was negligible, unlike the results for geometric patterns or letter strings. As n_{diff} increased, the $\overline{RT}$ decreased, and at a faster rate with larger ISI. The structure of the data for $\overline{RT}_{\text{same}}$ depended on the ISI: for ISI = 1 sec, $\overline{RT}_{\text{same}}$ was indeed relatively unaffected by n_{rel}, but with intervals of 4 and 10 sec, $\overline{RT}_{\text{same}}$ increased with n_{rel}. Both $\overline{RT}_{\text{same}}$ and $\overline{RT}_{\text{diff}}$ were substantially greater than those in Bamber's experiment (1969). For example, with $n_{\text{rel}} = 3$ and $n_{\text{diff}} = 1$, values of $\overline{RT}_{\text{diff}}$ were approximately 1,050, 1,250, and 1,400 ms for ISIs of 1, 4, and 10 sec, respectively, and with $n_{\text{rel}} = 3$, values of $\overline{RT}_{\text{same}}$ were approximately 1,050, 1,350, and 1,550 ms, respectively, for the three ISIs. Consider what these data suggest about P_{same} and P_{diff} at short and long ISIs. Do they differ? Is either holistic? Is either parallel? Is P_{diff} self-terminating? How might the increase in ISI change the representation being compared? How might it change the process or processes of comparison? What, if anything, do these results say about brain mechanisms? Compare the overall $\overline{RT}$s to those in figures 9.3 and 9.4. What might the difference mean?[42]

9.6 *Issues of experimental design.* We have seen that variation of the number of relevant attributes, n_{rel}, has provided useful information for distinguishing among theories. There are at least four ways in which n_{rel} could be varied. In method 1, the values of irrelevant attributes are held constant from stimulus to stimulus and trial to trial. In method 2, values of the irrelevant attributes are permitted to vary between trials, but do not differ between the two

stimuli on the same trial. In method 3, values of attributes that are irrelevant vary in the same way as when they are relevant; only the instructions to the subject and the mapping of stimulus pairs onto correct responses R_{same} and R_{diff} differ as n_{rel} is changed. Method 4 is like method 3, except that the numbers of "same" and "different" trials are adjusted as n_{rel} is varied so the proportions of these two trial types remain constant. Method 1 was used by Hawkins (1969), for example, while method 4 was used by Egeth (1966). Compare and contrast the four methods in terms of which other factors will vary as a consequence of the experimenter's manipulation of n_{rel}, what effects on performance such "confounded" variation might have, and how these effects might bear on the inferences we can draw from the data. Among the issues you might consider are how much subjects must remember as they perform the task, how many attributes they are likely to encode, how much ignoring ("filtering") of stimulus differences they must do, and the relative frequency of the two responses.

9.7 *The shortest* RT_{diff}. In this chapter the only aspect of the RT data from a condition we have considered is the mean, except for a brief mention of the standard deviation in section 9.2.7.4. But RTs from the same condition take on different values from trial to trial: they have a distribution. Increasingly, we are discovering that other aspects of the distribution, in addition to the mean, have important things to say about the underlying mechanism, and that the hypothetical mechanisms we consider have interesting predictions to make about other aspects of the distribution. As an example of such a prediction, consider the sequential mechanism for R_{diff}, assuming that the search path is random and that the encoding process duration does not increase with n_{rel}. What happens as n_{rel} is increased from 1 to 3 while $n_{diff} = 1$? Let RT_1 and RT_3 denote the corresponding sets of RTs.

When $n_{rel} = 1$, the first test will be a mismatch, so that on all trials the RT will contain the duration of no matching tests: $n_{mtests} = 0$. We can write $RT_{n_{rel}}(n_{mtests}) = RT_1(0)$ for the set of times. When $n_{rel} = 3$, there are three possibilities that will occur with equal probability: either the first, second, or third test will be a mismatch; n_{mtests} will be 0, 1, or 2. As experimenters, we will not know which it will be for any particular trial, but (given the model) we will know that the three possibilities are equally likely. RT_3 will thus be an equal-probability mixture of $RT_3(0)$, $RT_3(1)$, and $RT_3(2)$. The means of these three component sets of RTs will increase as n_{mtests} increases.

If $n_{mtests} = 0$, then the sequence of operations that determines the RT is the same, whether n_{rel} is 1 or 3. More generally, if we know n_{mtests} for a particular trial, then we can learn nothing further about the RT for that trial by also knowing n_{rel}. We can therefore drop the subscript on $RT_{n_{rel}}(n_{mtests})$ and say, simply, that $RT_1 = RT(0)$, and RT_3 is an equal-probability mixture of $RT(0)$, $RT(1)$, and $RT(2)$. Thus one of the three components of the RT_3 mixture is indistinguishable from RT_1. Consider the relation between the shortest RT observed when $n_{rel} = 1$ versus 3. We have to worry about the number of trials in each condition because the shortest observed value will, in general, decrease as the sample size increases (statistical facilitation again). Suppose therefore that we have 50 observations for $n_{rel} = 1$, which we can call obsns(1, 50), and 150 observations for $n_{rel} = 3$, which we can call obsns(3, 150), and consider the expected relationship between min {obsns(1, 50)} and min {obsns(3, 150)}. Among obsns(3, 150) will be about 50 from $RT(0)$. If values in the $RT(1)$ and $RT(2)$ sets are much larger than values in the $RT(0)$ set, then min {obsns(3, 150)} will come from the $RT(0)$ set, and will on average be as small as min {obsns(1, 50)}. If values in the other two sets are not much larger, then having them mixed in with the $RT(0)$ set can only make min {obsns(3, 150)} still smaller. It follows that even though $\overline{RT}_3 > \overline{RT}_1$, we expect that min {obsns(3, 150)} $\leq$ min {obsns(1, 50)}. Thus the fastest trials in a hard condition are likely to be as fast as the fastest trials in an easy condition. This property of self-terminating search is an example of one that increases the power of the set of tools available for model testing. What might happen to this relationship if, despite our assumption to the contrary, encoding time increased with n_{rel}?

9.8 *Inferring the duration of a computer operation by the subtraction method.* One way for a computer to produce a time interval T is to start a clock at zero and read it periodically until its value, "clockval," equals or exceeds T. This could be done in the C language by executing the loop:

(1) *while* (*clockval* < T) *clockread*(*&clockval*);

where "&clockval" specifies the address into which the clock value should be placed each time the clock is read. We recently had to estimate the time, α^*, from one clock-read to the next when our lab computer executed this loop. One way to measure α^* would be to determine how many clock-reads occurred during time T.[43] The problem is that command (1) does not provide this information. To count clock-reads, we had to elaborate the command by concatenating an indexing operation, $i{+}{+}$, with the clock-read

(2) *while* (*clockval* < T) {$i{+}{+}$; *clockread*(*&clockval*);},

where $i{+}{+}$ means "i becomes $i + 1$." In the initialization we set $i = 0$. Now we could count how many times the clock had been read during time T. Let this number be $N_1(T)$, where the subscript tells us that one indexing operation is included in the loop. The problem is that the indexing operation adds an unknown time increment, β^*, to the desired α^*. This situation is analogous to one described in appendix 2 on the subtraction method. The desired loop duration, α^*, is like the duration, α, of residual mental operations. The indexing duration, β^*, is like the duration, β, of a matching letter test. Just as there seems to be no experiment that provides a good direct measure of α (because encoding is likely to be different when no comparison is required), so we could not find a way to directly measure the time per iteration in command (1) (because that command provides no count). On the other hand, with indexing we could measure $\alpha^* + \beta^*$, which is not what we wanted to know. To solve this problem, we also collected data on the performance of the same command, but with a second $i{+}{+}$ indexing operation concatenated with the first and the clock-read:

(3) *while* (*clockval* < T) {$i{+}{+}$; $i{+}{+}$; *clockread*(*&clockval*);}.

This would provide us with $N_2(T)$, and a measure of $\alpha^* + 2\beta^*$. For a validity check we also determined $N_3(T)$, and, for a better test of linearity, $N_4(T)$ as well. We set T at 1.3 sec and obtained the following values of the four counts: $N_1(1.3) = 89{,}551$, $N_2(1.3) = 78{,}357$, $N_3(1.3) = 69{,}453$, and $N_4(1.4) = 62{,}441$. Show that $\hat{\alpha}^*_{\text{extrap}} = 12.4$ microseconds (the desired measurement of the time between clock-reads), that $\hat{\beta}^* = 2.1$ microseconds (the duration of one indexing operation), and that the validity check was successful. Are you convinced?

9.9 *Serial-position effects.* Suppose the search path in Bamber's experiment (1969) was consistently left to right across the letter string, and you plotted $\overline{RT}_{\text{diff}}$ for $n_{\text{diff}} = 1$ as a function of the serial position of the mismatching letter, separately for $n_{\text{rel}} = 2$, 3, and 4. How should these functions look? Suppose, instead, the search path was random from trial to trial.

Notes

I am grateful to Janice Hamer, Michael Kahana, Teresa Pantzer, Seth Roberts, Don Scarborough, and Jennifer Sternberg for their very helpful comments on earlier drafts.

1. A bar over a variable denotes the mean (or expectation) of that variable.
2. Limiting the number of values to two might be an error because it is probably a smaller number than is typically encountered in real life. We would like to learn about what people do "naturally," rather than about special mental strategies they might develop for particular laboratory tasks. We can draw some reassurance from the fact that experiments with two-valued and three-valued attributes give similar results; and two-valued examples serve well for illustration.

3. See, for example, Van Essen, Anderson, and Felleman 1992.
4. See appendix 1 for a discussion of the role of error rate in the interpretation of RT data.
5. The first string (the "target string") was viewed ad lib; the second string (the "test string") was displayed briefly, for 100 ms.
6. Any letter that appeared in both strings appeared in the same position; subjects never saw pairs such as **KSV, KVS**, or **KSV, VTK**.
7. A proof of equation 9.2 is sketched in question 9.4 at the end of the chapter.
8. When the level of one factor modulates the effect of changing the level of another, their effects are said to "interact." (See chaps. 12 and 14, this volume.)
9. Results from some kinds of memory search experiment violate the slope ratio, negative slope: positive slope = 2 : 1, which is diagnostic of a self-terminating testing process, and instead show linear functions with a 1 : 1 slope ratio, which some researchers have interpreted as indicating "exhaustive search"—that is, a process in which all items are tested on both positive and negative trials. (See, for example, Sternberg 1975.)
10. If the two patterns are presented simultaneously, then the duration of the encoding operations for both patterns is incorporated in the RT. This is perhaps one reason why $\overline{RT}$ in the simultaneous condition is longer than $\overline{RT}$ in the serial condition, as shown in figure 9.3A by the curves labeled "SIM" and "SER." Note, however, that the amount by which $\overline{RT}$ is shortened by the successive versus simultaneous display is greater for R_{same} than for R_{diff}, which may complicate the interpretation of the shortening.
11. Greek letters such as α represent constants or parameters in the models.
12. In chapter 14, this volume, on the additive-factor method (AFM), I discuss how to test assumptions of stages and selective influence.
13. Without constraint 3, constraint 4 would have to be expanded into two separate statements, one for tests of different attributes leading to a match, and one for tests of different attributes leading to a mismatch.
14. Nor does it matter whether the durations of successive processes are correlated in some way rather than being independent, or what the forms of their duration distributions are. (See Wickens, chap. 12, this volume, for the idea of a distribution.) This is not to say that such characteristics are in general irrelevant. For example, they become important if we wish to understand the effects of experimental factors on RT variability as well as on $\overline{RT}$.
15. See, for example, figures 4D and 4E in Sternberg 1969.
16. In the present context, where the experimenter's goal is for error rates to be low, it is reasonable to use the speed of same-different decisions (rather than a more traditional accuracy measurement) to define ease of discrimination. Many investigators would expect that if the values of one attribute or letter position were less discriminable than another, this would be reflected in both $\overline{RT}_{\text{same}}$ and $\overline{RT}_{\text{diff}}$. It follows that if $\overline{RT}_{\text{diff}}$ were approximately equated across attributes (or letter positions), then $\overline{RT}_{\text{same}}$ would also be equated. Likewise, if attributes were ordered according to either $\overline{RT}_{\text{same}}$ or $\overline{RT}_{\text{diff}}$, they would then have the same ordering with respect to the other. However, these assumptions have not been tested, to my knowledge.
17. Of course, if different subjects have different fixed paths, then the mean over subjects would represent a mixture, and the mean data might be a poor reflection of individual behavior.
18. Others might say the theory is "too powerful" because it can "explain" too much. That is, by appropriate choice of values for the β and γ parameters, the theory could be made to explain many alternative data patterns.
19. I was first impressed with such theoretical flexibility in my doctoral research on two-choice learning, where I worked with four models that differed greatly in the effect of the response produced on trial m on the response probability on trial $m + n$, ranging

from no effect to an effect that continued, undamped, for all n. All four models could explain the learning curve, and only one of the four could not also explain the distribution of the lengths of runs of errors. This forced me to search for other properties that would discriminate among the four models (Sternberg 1963, section 4.5).

20. A numerical example of this phenomenon under more complicated conditions—where γ_A, γ_S, and γ_D have overlapping distributions as in the present example, but also unequal means—is shown in table 9.5 and discussed in section 9.2.7.3.
21. As mentioned in question 9.5 at the end of the chapter, this was not found in an experiment on face recognition, where the features were mouth, nose, and so on (Smith and Nielsen 1970). This suggests that for faces, R_{diff} is based on parallel or holistic tests, rather than sequential ones.
22. These properties of invariance and additivity of the effects of factors that influence different operations, when the operations are arranged as stages, are elaborated in chapter 14, this volume, and form the basis of the method of additive factors discussed in that chapter.
23. An alternative measure of "goodness of fit," used by Massaro in chapter 8, this volume, and minimized by the "least squares" fitting procedure, is the square root of the mean squared deviation (RMSD), which is 3.5 and 4.9 ms, respectively, for panels A and B.
24. In a more searching comparative evaluation of models, the fitting would probably be done separately for stable data from practiced individual subjects, and statistical tests (see Wickens, chap. 12, this volume) would be an important part of the evaluation.
25. In examining the effects of n_{diff} on $\overline{RT}_{diff}$, we would typically compare the mean of the $n_{diff} = 2$ case for the feature pair A, S (138 ms) to the mean of the $n_{diff} = 1$ cases for the same features, A and S, taken individually (175 ms); this avoids confusing effects of n_{diff} with differences of mismatch durations from one feature to another.
26. That an inverse relation between n_{diff} and $\bar{T}_{ft}$ with diminishing returns occurs with the particular distributions of test durations assumed in the example of table 9.5 does not prove that it is always true. Thus the example shows that a parallel testing process *can* produce the phenomenon, not that it *must* produce it. However, I believe that it must produce it, for plausible test-duration distributions.
27. This example illustrates the fact that the mean of a sum of random variables is the sum of their means, whatever their distributions and whatever their correlation.
28. Calculations such as these are better applied to the data from individual subjects. If there are individual differences, then ratios of means of subject data may misrepresent the individual subject ratios.
29. As we have seen, the parallel-test model requires these differences not to be influenced by n_{rel}; we expect, for example, that $\overline{RT}_{41} - \overline{RT}_{42} = \overline{RT}_{31} - \overline{RT}_{32} = \overline{RT}_{21} - \overline{RT}_{22}$. For this model, therefore, the observed values of each of these differences estimates the effect of increasing n_{diff} from 1 to 2; an estimate of the one-step reduction from $n_{diff} = 1$ to 2 that used all of the available information in the data would thus be the mean of the three differences. The systematic differences among the three, already noted, which in one context provide evidence against the parallel-test model, must be regarded as "noise" in getting estimates of properties of this model, because when we obtain the estimates we operate as if the model were "true."
30. See Wickens, chapter 12, this volume.
31. A distribution of durations is often conveniently specified by the proportion of durations no greater than τ, $\Pr\{T \leq \tau\}$ over the range of τ. For the exponential distribution, the range is $\tau \geq 0$, and $\Pr\{T \leq \tau\} = 1 - e^{-\tau}$.
32. For a rectangular (continuous uniform) distribution, the proportion of occurrences of each value is a constant over some range of values (say, from 40 to 80 ms) and zero elsewhere.

33. Note the distinction between one relevant feature that mismatches, (**F**), and two relevant features, one of which mismatches, (**F**, *s*).
34. One such demonstration would be to show that we can induce a change in the serial-position effect without altering the mean effects of n_{diff} and n_{rel}. This would support the sequential-test model.
35. Thus as we introduce more parallel tests, the same (exponential) distribution (illustrated by the bottom curve in figure 9.9C) is associated with strongly diminishing returns for the minimum, which is required by $\overline{RT}_{diff}$, and with weakly diminishing returns for the maximum, which is required to lessen the model-data disparity for $\overline{RT}_{same}$.
36. It is interesting to consider what this possibility suggests about the ranges of RT (shortest and longest values) for R_{same} versus R_{diff}.
37. Although we might initially be tempted to think of this simple possibility as a necessary consequence of the $P_{same} \rightarrow P_{diff}$ structure ("If P_{same} occurs first, and the factor influences P_{same}, then because RT_{diff} includes the duration of both P_{same} and P_{diff}, the factor must influence RT_{diff} as well as RT_{same}."), there are at least two reasons why it is not. First, the factor might influence a component of P_{same} that occurs only on "same" trials, such as the decision process associated with R_{same}; in that case, there would be no reason to expect an effect on $\overline{RT}_{diff}$. And second, the factor might influence P_{diff} as well as P_{same}; in that case, we would expect different size effects of the factor on $\overline{RT}_{same}$ and $\overline{RT}_{diff}$.
38. Such holistic comparison of visual patterns is sometimes described as "template comparison" (e.g., Smith and Nielsen 1970; or see Massaro, chap. 8, this volume). But template comparison can also be regarded as an analytic feature-testing process, where each pixel in the test stimulus is compared to the pixel in the corresponding position in the template; a feature is then the darkness or lightness (or the color) of one pixel.
39. See question 9.5 at the end of the chapter for more information about this experiment, which suggests that the face-comparison process may differ qualitatively from the comparison of geometric patterns or letter strings.
40. It would be unwise, for example, to design the experiment so that n_{rel} increased systematically from 1 to 4 as the subject became increasingly practiced, because this would produce a confounding between n_{rel} and practice. Instead, Bamber (1969) arranged for n_{rel} to vary randomly from trial to trial.
41. One likely source of this idea is the relative operating characteristic (ROC) of signal detection theory (SDT), discussed by Swets (chap. 13, this volume), which describes the trade-off between the frequencies of two kinds of error. Because SDT is applied to experiments with two kinds of trials (call them "positive" and "negative"), two alternative responses (positive and negative), and sufficient inherent uncertainty, substantial numbers of errors are likely to occur on both kinds of trials (false negatives on positive trials, and false positives on negative trials). Under these conditions, subjects cannot escape the need to adopt a position on the ROC, and subjects appear to be able to move freely along it. It is an open question whether such arbitrariness of strategy also applies to the corresponding "speed-accuracy operating characteristic" in an RT experiment, where something close to perfect accuracy is an option.
42. It should be mentioned that some of the data trends in Smith and Nielsen 1970, though substantial, were reported not to be statistically significant. Also, because n_{rel} was fixed for a block of trials, and for a block with a given n_{rel}, $\min(n_{diff}) = n_{rel} - 3$, the average and minimum levels of discriminability on R_{diff} trials increased markedly as n_{rel} increased. This confounding of the discriminability of stimuli in a block with the n_{rel} value for that block might have partially counteracted the effect of n_{rel}.
43. Other solutions may occur to you; given our system, those that occurred to us were either at least as complex, or not feasible.

References

Ashby, F. G., and Maddox, W. T. (1994). A response time theory of separability and integrality in speeded classification. *Journal of Mathematical Psychology* 38, 423–466.

Baker, N. (1988). *The mezzanine.* New York: Weidenfeld and Nicholson.

Bamber, D. (1969). Reaction times and error rates for "same"-"different" judgments of multidimensional stimuli. *Perception & Psychophysics* 6, 169–174.

Bamber, D. (1972). Reaction times and error rates for judging nominal identity of letter strings. *Perception & Psychophysics* 12, 321–326.

Calvino, I. (1981). *If on a winter's night a traveler.* Orlando, FL: Harcourt Brace Jovanovich.

Cattell, J. M. (1947). *James McKeen Cattell, 1860–1944: Man of science.* Lancaster, Science Press.

Chamberlin, T. C. (1890). The method of multiple working hypotheses. *Science* (old series) 15, 92–96. Reprinted (1965) in *Science* 148, 754–759.

Chase, W. G. (1978). Elementary information processes. In W. G. Estes, *Handbook of learning and cognitive processes.* Vol. 5, *Human information processing,* pp. 19–90. Hillsdale, NJ: Erlbaum.

Corcoran, D. W. J. (1971). *Pattern recognition.* Harmondsworth, England: Penguin.

David, F. N. (1970). *Order statistics.* New York: Wiley.

Donders, F. C. (1868). On the speed of mental processes. Translated from the Dutch by W. G. Koster. In Koster (Ed.), *Attention and performance II. Acta Psychologica* 30 (1969), 412–431.

Doucet, P., and Sloep, P. B. (1992). *Mathematical modeling in the life sciences.* Chichester: Ellis Horwood.

Egeth, H. (1966). Parallel versus serial processes in multidimensional stimulus discrimination. *Perception & Psychophysics* 1, 245–252.

Egeth, H., and Blecker, D. (1971). Differential effects of familiarity on judgments of sameness and difference. *Perception & Psychophysics* 9, 321–326.

Eichelman, W. H. (1970). Familiarity effects in the simultaneous matching task. *Journal of Experimental Psychology* 86, 275–282.

Farah, M. J. (1992). Is an object an object an object? Cognitive and neuropsychological investigations of domain specificity in visual object recognition. *Current Directions in Psychological Science* 1, 164–169.

Farell, B. (1985). "Same"-"different" judgments: A review of current controversies in perceptual comparisons. *Psychological Bulletin* 98, 419–456.

Farell, B. (1988). Comparison requirements and attention in identical-nonidentical discrimination. *Journal of Experimental Psychology: Human Perception and Performance* 14, 707–715.

Feller, W. (1968). *An introduction to probability theory and its applications.* Vol. 1. 3d ed. New York: Wiley.

Giray, M., and Ulrich, R. (1993). Motor coactivation revealed by response force in divided and focused attention. *Journal of Experimental Psychology: Human Perception and Performance* 19, 1278–1291.

Hawkins, H. L. (1969). Parallel processing in complex visual discrimination. *Perception & Psychophysics* 5, 56–64.

Howson, C. (1990). Fitting your theory to the facts: Probably not such a bad thing after all. In C. W. Savage (Ed.), *Minnesota studies in the philosophy of science.* Vol. 14, *Scientific theories,* pp. 224–244. Minneapolis: University of Minnesota Press.

Howson, C., and Urbach, P. (1993). *Scientific reasoning: The Bayesian approach.* 2d ed. La Salle, IL.: Open Court.

Jastrow, J. (1890). *The time-relations of mental phenomena.* Fact and Theory Papers, no. 6. New York: Hodges.

Keele, S. W. (1986). Motor control. In K. R. Boff, L. Kaufman, and J. P. Thomas (Eds.), *Handbook of perception and human performance.* Vol. 2, *Cognitive processes and performance.* pp. 30-1–30-60. New York: Wiley.

Koster, W. G., Ed. (1969). *Attention and performance II. Acta Psychologica* 30.

Kreuger, L. (1978). A theory of perceptual matching. *Psychological Review* 85, 278–304.

Lockhead, G. (1972). Processing dimensional stimuli: A note. *Psychological Review* 79, 410–419.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization.* New York: Oxford University Press.

McElree, B., and Dosher, B. A. (1989). Serial position and set size in short-term memory: The time course of recognition. *Journal of Experimental Psychology: General* 118, 346–373.

Meijers, L. M. M., and Eijkman, E. G. J. (1977). Distributions of simple RT with single and double stimuli. *Perception & Psychophysics* 22, 41–48.

Meyer, D. E., Irwin, D. E., Osman, A. M., and Kunios, J. (1988). The dynamics of cognition and action: Mental processes inferred from speed-accuracy decomposition. *Psychological Review* 95, 183–237.

Meyer, D. E., Osman, A. M., Irwin, D. E., and Yantis, S. (1988). Modern mental chronometry. *Biological Psychology* 26, 3–67.

Miller, J. (1978). Multidimensional same-different judgements: Evidence against independent comparison of dimensions. *Journal of Experimental Psychology: Human Perception and Performance* 4, 411–422.

Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology* 14, 247–279.

Miller, J., and Bauer, D. W. (1981). Irrelevant differences in the "same"-"different" task. *Journal of Experimental Psychology: Human Perception and Performance* 7, 196–207.

Nickerson, R. S. (1967). "Same"-"different" response times with multi-attribute stimulus differences. *Perceptual and Motor Skills* 24, 543–554.

Nickerson, R. S. (1972). Binary-classification reaction time: A review of some studies of human information-processing capabilities. *Psychonomic Monograph Supplements* 4, 275–318.

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology* 43, 25–53.

Pachella, R. G. (1974). The interpretation of reaction time in information-processing research. In B. H. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition,* pp. 41–82. Hillsdale, NJ: Erlbaum.

Posner, M. I., and McLeod, P. (1982). Information processing models: In search of elementary operations. *Annual Review of Psychology* 33, 477–514.

Proctor, R. W. (1981). A unified theory for matching-task phenomena. *Psychological Review* 88, 291–326.

Proctor, R. W., Rao, K. V., and Hurst, P. W. (1984). An examination of response bias in multiletter matching. *Perception & Psychophysics* 35, 464–476.

Raab, D. H. (1962). Statistical facilitation of simple reaction times. *Transactions of the New York Academy of Sciences* 24, 574–590.

Ratcliff, R., and Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review* 83, 190–214.

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

Roberts, S., and Sternberg, S. (1993). The meaning of additive reaction-time effects: Tests of three alternatives. In D. E. Meyer and S. Kornblum (Eds.), *Attention and performance*

XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience, pp. 611–653. Cambridge, MA: MIT Press.

Schweickert, R. (1985). Separable effects of factors on speed and accuracy: Memory scanning, lexical decision, and choice tasks. *Psychological Bulletin* 97, 530–546.

Schweickert, R. (1993). Information, time, and the structure of mental events: A twenty-five year review. In D. E. Meyer and S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience*, pp. 535–566. Cambridge, MA: MIT Press.

Sergent, J., and Takane, Y. (1987). Structures in two-choice reaction-time data. *Journal of Experimental Psychology: Human Perception and Performance* 13, 300–315.

Shaw, M. L. (1982). Attending to multiple sources of information: 1. The integration of information in decision making. *Cognitive Psychology* 14, 353–409.

Smith, E. E. (1968). Choice reaction time: An analysis of the major theoretical positions. *Psychological Bulletin* 69, 77–110.

Smith, E. E., and Nielsen, G. D. (1970). Representations and retrieval processes in short-term memory: Recognition and recall of faces. *Journal of Experimental Psychology* 85, 397–405.

Sternberg, S. (1963). Stochastic learning theory. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.), *Handbook of mathematical psychology*. Vol. 2, pp. 1–120. New York: Wiley.

Sternberg, S. (1966). High-speed scanning in human memory. *Science* 153, 652–654.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. In W. G. Koster, (Ed.), *Attention and performance II. Acta Psychologica* 30, 276–315.

Sternberg, S. (1975). Memory scanning: New findings and current controversies. *Quarterly Journal of Experimental Psychology* 27, 1–32.

Townsend, J. T. (1990). Serial vs. parallel processing: Sometimes they look like Tweedledum and Tweedledee but they can (and should) be distinguished. *Psychological Science* 1, 46–54.

Townsend, J. T., and Ashby, F. G. (1983). *The stochastic modeling of elementary psychological processes*. Cambridge: Cambridge University Press.

Van Essen, D. C., Anderson, C. H., and Felleman, D. J. (1992). Information processing in the primate visual system: An integrated systems perspective. *Science* 255, 419–423.

Vorberg, D. (1981). Reaction time distributions predicted by serial self-terminating models of memory search. In S. Grossberg (Ed.), *Symposium in applied mathematics*. Vol. 13, *Mathematical psychology and psychophysiology*, pp. 301–318. Providence, RI: American Mathematical Society.

Welford, A. T., Ed. (1980). *Reaction times*. London: Academic Press.

Wickelgren, W. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica* 41, 67–85.

Woodworth, R. S. (1938). *Experimental psychology*. New York: Holt.

Yantis, S. G., Meyer, D. E., and Smith, J. E. K. (1991). Analyses of multinomial mixture distributions: New tests for stochastic models of cognition and action. *Psychological Bulletin* 110, 350–374.