Psychonomic Bulletin & Review
1995, 2 (1), 20-54

Statistical mimicking of reaction time data: Single-process models, parameter variability, and mixtures

TRISHA VAN ZANDT
Johns Hopkins University, Baltimore, Maryland

and

ROGER RATCLIFF
Northwestern University, Evanston, Illinois

Statistical mimicking issues involving reaction time measures are introduced and discussed in this article. Often, discussions of mimicking have concerned the question of the serial versus parallel processing of inputs to the cognitive system. We will demonstrate that there are several alternative structures that mimic various existing models in the literature. In particular, single-process models have been neglected in this area. When parameter variability is incorporated into single-process models, resulting in discrete or continuous mixtures of reaction time distributions, the observed reaction time distribution alone is no longer as useful in allowing inferences to be made about the architecture of the process that produced it. Many of the issues are raised explicitly in examination of four different case studies of mimicking. Rather than casting a shadow over the use of quantitative methods in testing models of cognitive processes, these examples emphasize the importance of examining reaction time data armed with the tools of quantitative analysis, the importance of collecting data from the context of specific process models, and the importance of expanding the database to include other dependent measures.

Since the publication of Donders's (1868/1969) essay, "On the speed of mental processes," psychologists have measured the time required by experimental subjects to perform various tasks. These reaction times (RTs) and the changes in RT under different experimental manipulations have been used as evidence for or against models of mental architecture—the arrangement of the mental processes underlying the subject's performance (Sternberg, 1969; Townsend & Ashby, 1983; Woodworth, 1938, chapter 14). RT data have played an important role in distinguishing between models and in testing hypotheses about processes and structures. Consequently, considerable effort has been devoted to the refinement of RT measures, from techniques to optimize the accuracy of subsequent statistical analyses of RT summary statistics (Ratcliff, 1993; Townsend, 1990b; Ulrich & Miller, 1994) to the estimation of RT distributions and RT hazard functions (Burbeck & Luce, 1982; Luce, 1986; Ratcliff & Murdock, 1976). The concentration on the RT distributions can be seen as an advance from the use of less informative summary statistics such as the mean or median. In the distributional approach, the density or distribution functions predicted by various models are fit to the usually unimodal, positively skewed RT densities or distributions produced by experimental subjects (e.g., Green & Luce, 1971; Heathcote, Popiel, & Mewhort, 1991; Hockley, 1984; Hohle, 1965; McGill & Gibbon, 1965; Ratcliff, 1978, 1979, 1988; Ratcliff & Murdock, 1976). The success or failure of the fitting process indicates the appropriateness of a particular model for the task at hand.

In this paper, we wish to address the issue of statistical mimicking of RT data—that is, the ability of very different kinds of models to produce similar patterns of mean RTs and RT distributions. Highly dissimilar mental architectures often can produce RTs that are indistinguishable from each other, at least in the sense that appropriate statistical analyses applied to the data cannot determine any differences between the RT distributions. The existence of such statistical mimics to various models of performance raises concerns about the way that various tests proposed for RT analyses are performed, especially when these tests are applied without the constraints of processing models. In the first half of the paper, we will discuss several such tests and examine how they are able to distinguish between data generated by different kinds of models. In the second half of the paper, we will consider two models of RT performance that rely on multiple stages

This project was supported by NIMH Grants HD MH44640 and MH00871 to R.R. and was completed while the first author was a postdoctoral fellow at Northwestern University. The manuscript was greatly improved by helpful comments from Barbara Dosher, Ehtibar Dzhafarov, Thomas Fikes, Rich Schweickert, Saul Sternberg, and Jim Townsend. Correspondence may be addressed to T. Van Zandt, Department of Psychology, Ames Hall, Johns Hopkins University, 3400 N. Charles St., Baltimore, MD 21218-2686 (e-mail: trish@maigret.psy.jhu.edu).

of processing and demonstrate that a very different kind of model that does not rely on multiple processes can also fit the RT data. In so doing, we emphasize that analyses of RT alone are not sufficient to distinguish between these types of models. Additional measures, such as accuracy or confidence judgments and the observations of the behavior of RT distributions over a range of experimental conditions, are needed to determine the adequacy of these models of performance.

We must emphasize that the issue of model identifiability is not limited to the RT paradigms that we discuss here or to the area of cognitive psychology in general. This problem will arise across the different disciplines and is worst for those areas that have not benefited, as cognitive psychology has, from concentrated attempts to quantify psychological findings. Demonstrating that this issue is still a concern for cognitive psychology underscores the problem for those other areas. As cognitive psychologists, we work in the areas with which we are most familiar, but this should not be taken as a signal that other areas of experimental psychology are exempt from the issues we raise.

We will begin by outlining the mimicking problem and a solution that has been proposed to circumvent it. We will then discuss the issue of parameter variability, which undermines the utility of that solution. Later in the paper, these topics will be addressed concretely with four case studies, two concerning model free tests of processing and two concerning specific multiprocess models. In each of these case studies, we will present an alternative single-process model that either passes the test for a multiprocess model or accounts for RT data as well as the multiprocess model. Under no circumstances should these findings be interpreted as a demonstration of weakness in the tests. Rather, they demonstrate that there is a right way and a wrong way to apply them.

Mimicking

RT data have been collected for a wide variety of experimental paradigms. These data have been used to address questions concerning the serial or parallel operation of the processes involved in the task of interest, the nature of information transmission from one process to the next, and the hierarchical organization of the processes. The issue of serial versus parallel arrangement of the processes in memory retrieval initially received a great deal of attention (e.g., Sternberg, 1966; Townsend, 1972, 1974; Townsend & Ashby, 1983). This was due in part to the nature of the paradigms and stimulus materials used in memory "search" experiments. For instance, an experiment designed to test some hypothesis about memory function typically has subjects learn a list of words or other items and then presents them with a test item to which they should answer "old" or "new." A natural first question concerning the way the memory process operates is whether the list items in memory are compared with the test item serially (one at a time) or in parallel (simultaneously). Unfortunately, RTs do not readily distinguish between these two types of architectures. When a single independent variable is manipulated, such as list length, both serial and parallel models can produce identical patterns of RT. For example, for every model in which subprocesses operate in parallel and the time required to complete each subprocess has no influence on (is independent from) the amount of time required by any other subprocess, there exists a mathematically equivalent model, not discriminable from the parallel model, in which all subprocesses operate in series. The RTs produced by the independent parallel model and its equivalent serial representation will be identical in every way. Townsend's (1972) careful enunciation of this theoretical pitfall discouraged further research that relied on the premise that mean RT data alone could discriminate between serial and parallel processes in these kinds of tasks.

The existence of serial mimics to parallel processes is probably the most well-known example of an identifiability problem in cognitive modeling. The serial/parallel question itself is very specific, easily operationalized, and seemingly tractable, and so, at first blush, it appears to be just the type of question that cognitive psychology should devote itself to answering. Townsend (1990a) argues that the serial/parallel issue is indeed exactly the kind of problem that should be resolved, but, unfortunately, it has not been. Despite a growing body of theoretical work outlining how RT can be used to distinguish between serial and parallel processing (e.g., Roberts & Sternberg, 1994; Schweickert, 1980; Schweickert & Townsend, 1989; Sternberg, 1969; Townsend & Ashby, 1983; Townsend & Schweickert, 1989), empirical resolution of the serial/parallel issue seems to have fallen by the wayside. Some researchers might say that the reason for this is the question itself: since all physiological evidence suggests that processing is parallel at some level anyway, it might seem pointless to invest the effort required to debunk the serial "straw man." Also, the paradigms that addressed the question were somewhat limited, usually involving memory or visual search (although richer paradigms have since been proposed; cf. Schweickert & Townsend, 1989). The RT methodology is now applied to other, perhaps more interesting, questions for which the problem of model identifiability does not (yet) exist, such as attentional control (e.g., Treisman, Vieira, & Hayes, 1992; Wolfe, Cave, & Franzel, 1989), the nature of information flow through the system (e.g., McClelland,1979; Miller, 1988, 1993), the acquisition of skill (e.g., Carlson & Schneider, 1989; Logan, 1988, 1992), and so on. However, the way that RTs are employed to test hypotheses in other areas is often little changed from the way that they were applied to the serial/parallel question. Other performance variables are often ignored, and rigorously defined models of the processes of interest are often lacking. Therefore, these areas may also be prone to the problem of mimicking between various types of models.

The serial/parallel processing issue is one of exact mathematical equivalence of two models: the RTs that they predict can be identical, although the structures of

the models are very different. Even if exactly equivalent serial representations of independent parallel models did not exist, however, the RTs produced by the serial and parallel models might still be ambiguous. As several researchers have noted (e.g., Luce, 1986; Ratcliff, 1988), the unimodal and positively skewed RT density can easily be fit by a number of distributions. For example, the gamma, inverse-normal, and Ex-Gaussian distributions have all been shown to fit RT data to a greater or lesser degree (Ratcliff & Murdock, 1976). Because the shape of the distributions are highly similar, this "statistical mimicking" of RT data is still a concern even when mathematically equivalent relationships, such as those that arise in the serial/parallel case, do not exist. We now discuss issues of statistical power and the number of observations needed to discriminate between models that predict similar RT distributions, after which we will demonstrate how the hazard function can be used to provide finer discrimination between models and how parameter variability (leading to mixtures of distributions) reduces statistical power and the diagnosticity of the hazard function.

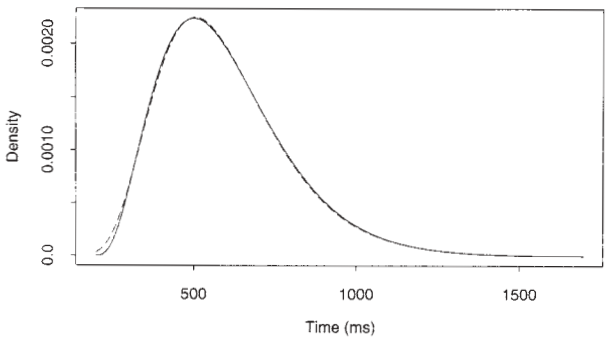

Figure 1. Statistical mimicking of a gamma density function with a rate of .01, a shape parameter of 4, and a base time parameter of 200 msec by an inverse-normal density with a mean of 597 msec and λ parameter of 5,244 msec (see Luce, 1986, p. 509).

Statistical Power

Consider the two density functions shown in Figure 1. It is very difficult to distinguish between them because they are so similar. Nonetheless, the two curves arise from very different functions (both have been proposed as models of the RT distribution). One density is a gamma (the solid line), and the other is an inverse normal (the dashed line). The gamma density might arise from a model in which several stages of processing must be completed, and each stage finishes with a time that is exponentially distributed. The inverse-normal distribution might arise from a diffusion model with a single response boundary. The analyses performed on RTs produced by the multiple-process model (the gamma) would not permit the experimenter to rule against a single-process model in which the RTs were inverse-normally distributed, because there are not enough differences between the two distributions. It might be argued that this is simply a problem of statistical power; with a sufficient number of RTs, differences between the gamma and the inverse normal will become evident. This is true: statistical mimicking is in part a problem of statistical power. If there is enough power in the test or comparison to be performed, then different hypotheses may be discriminated and the problem of mimicking is a problem no longer. At what point, however, does a lack of statistical power become an insurmountable obstacle, resulting in a problem of identifiability equivalent to that observed when two different models are mathematically indistinguishable?

Consider again the densities in Figure 1. We simulated eight sets of data, each set containing a number of observations from both distributions shown. We then calculated the asymptotic Kolmogorov-Smirnov and Kuiper statistics for each data set. (Both the Kolmogorov-Smirnov and Kuiper tests are nonparametric tests of the hypothesis that two independent samples were drawn from the same population.) The results of these analyses are presented in Table 1, along with the sample sizes from each set. Not until the sample sizes exceeded 40,000 observations from each distribution did the small differences between the distributions become statistically reliable. This suggests that, if we assume that a subject is able to perform 48 trials per minute in a 1-h session (1,250 msec for the total trial time including the intertrial interval, and no rest breaks allowed within the session—a tall order), and the subject's RTs were generated by a process producing gamma-distributed RTs, an experimenter would require at least 15 sessions from the subject to be able to reject the hypothesis that the RTs were generated by a process with inverse-normal finishing times. This also assumes that the parameters of the process remained constant across all trials and sessions: no fatigue, practice, or time-of-day effects, for example, and no effect of differing stimuli.

The only study of which we are aware that even approached this magnitude is that of Green and Luce (1971), who collected over 10,000 observations from their subjects. The study resolving the stage versus random walk issue above would be at least four times larger. It must be emphasized again that each of the resulting 40,000 observations must be identically distributed. The stimulus must be the same on every trial, the data must be free of any repetition effects, the subject's attention cannot at any point waver from the task, and so on. If it is believed that these requirements can be fulfilled, then the issue of statistical mimicking is irrelevant. However, while we agree that it might be theoretically possible to avoid the problem of statistical mimicking by simply collecting enough data points, in practice it may not be feasible.

To attempt to circumvent the necessity of collecting extraordinarily large numbers of observations, many researchers have advocated the use of the hazard function (the ratio of the density to one minus the distribution function) to discriminate between different distributions (e.g., Balakrishnan & Ashby, 1992; Bloxom, 1984, 1985; Burbeck & Luce, 1982; Luce, 1986). The hazard function gives the likelihood that, for a particular point in

Table 1
Results of the Nonparametric Kolmogorov-Smirnov and Kuiper Tests to Determine the Discriminability of the Two Variables Pictured in Figure 1

	Statistic	
Sample Size	Kolmogorov-Smirnov	Kuiper
100	.495	.919
1,000	.872	1.342
10,000	.792	1.478
20,000	1.145	1.410
30,000	1.196	1.474
40,000	1.389*	1.630
50,000	1.755†	2.008†
60,000	1.897‡	2.275‡

*$p < .05$. †$p < .01$. ‡$p < .005$.

time, an event occurs given that it has not occurred up to that point. Luce (1986) presented a demonstration of how different the shapes of hazard functions of different random variables can be, even though their density functions are very similar. Thus, the shapes of the hazard functions might be a useful way to discriminate between the inverse-normal and gamma densities in Figure 1. The hazard functions of the gamma and inverse-normal densities pictured in Figure 1 are shown in Figure 2. Although the hazard functions of these variables are also quite similar, the divergence in the tails of the functions is much clearer than the small differences between the densities. Note that the asymptotic behavior of the hazard functions is usually very difficult to observe, because it is determined by the last few percentage points of the data. Therefore, the tail of an empirical hazard function is generally quite unstable. The number of observations required to make the estimates of these tails stable is another question that we will not address here (but see Bloxom, 1984, 1985).

A Limitation of the Hazard Function

Consider the following example of statistical mimicking. In a diffusion model (e.g., Ratcliff, 1978) of response selection, a central processor keeps track of the level of "evidence" growing, or *drifting*, toward alternative responses. On average, depending on a particular trial, the evidence drifts in either a positive or a negative direction. A response is selected when the level of evidence is sufficiently positive or sufficiently negative, and this event is represented by the evidence level crossing a boundary. Usually there are two boundaries, one positive and one negative, corresponding to the two response alternatives in a two-choice task. Ratcliff (1988) has demonstrated that the diffusion model can closely mimic RT results produced by a multiple-subprocess model, in which the manipulation of some experimental variable results in the insertion of an additional serial subprocess with exponentially distributed processing time. (We will discuss this model in some detail in the following section.) The diffusion model was fit to experimental data, and the finishing times of the diffusion process were found to mimic the slowing of a serial process by the additional exponentially distributed subprocess. The diffusion RTs were not distributed exactly as the serial multiple-subprocess RTs, but the difference failed to exceed statistical criteria. Thus, in a statistical sense, the two models could not be distinguished.

The next question is whether or not hazard functions of the diffusion model and the serial model can be used to discriminate between the two (see, e.g., the gamma and inverse-normal distribution case presented earlier). The answer to this question is partly an empirical one. Most observed RT hazard functions are nonmonotonic: they increase and then decrease (e.g., Burbeck & Luce, 1982). Note that if the decision boundaries of the diffusion model are asymmetric, the processing-time distribution (defined by the time at which the diffusion process crosses the boundary closest to the starting point) for a given drift rate tends to the inverse normal as the asymmetry increases. The hazard function of the inverse-normal distribution is increasing to asymptote or increasing then decreasing over time, as are most empirical RT hazard functions. The hazard function of the serial model may assume a number of shapes, and additional assumptions would need to be made about the processes to which the serial exponential stage was added in order to know more. Thus, both models might accommodate the observed RT hazard functions.

Another reason that the hazard functions might not be diagnostic in this case is because the diffusion model incorporates a variable parameter, the drift rate, which is normally distributed. The variable drift rate has the effect of producing finishing-time distributions that are actually mixtures of observations arising from many different distributions, each conditioned on a different value of the drift rate.[1] The hazard functions of mixture distributions can look very different from the hazard functions of the individual distributions composing the mixture. In particular, mixture hazard functions can be nonmonotonic, even if the component hazard functions are, say, monotonic increasing (see, e.g., Barlow & Proschan, 1975). Only when the component hazard functions are constant or decreasing (nonincreasing) is the shape of the mixture hazard function constrained, and then it must also be nonincreasing. Even if the boundaries of the diffusion process are not asymmetric, the diffusion model

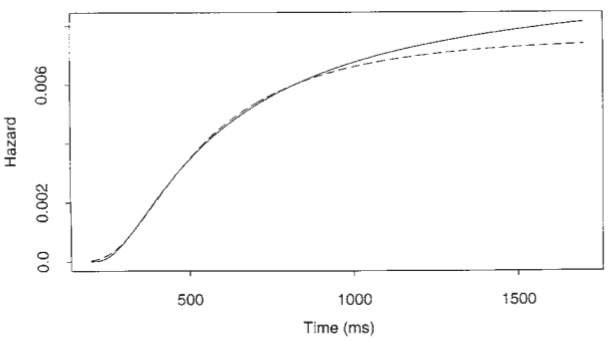

Figure 2. Hazard functions of the gamma (solid line) and inverse-normal (dashed line) variables shown in Figure 1.

predicts nonmonotonic hazard functions similar to the empirical hazard functions. We will discuss this issue in more detail later, but the main point is that parameter variability, which is likely to be present to some extent in any experimental situation, can rob the hazard function of its diagnosticity, compounding the problem of statistical mimicking of RT distributions.

This example also demonstrates that the class of single-process models, of which the single-process diffusion model is a member, has been relatively neglected when issues of mimicking among multiple-process models are considered. Clearly, some single-process models can mimic the RT distributions produced by multiple-process models, at least statistically. The additional concern of parameter variability makes the task of discriminating between the two types of models even more difficult. We therefore wish to emphasize the importance of attacking the mimicking problem from within specific process models, both single and multiple process, and with the added information provided by dependent measures other than RT. It is not easy to find a model that captures all aspects of the RT data, but there are ways to test between different competing models, evaluating their strengths and weaknesses and evaluating their performance with respect to a wide range of behavioral variables. Post hoc fitting of various probability distributions to observed RT data is useful as a way to summarize data; however, without the benefit of a more comprehensive process that encompasses other behavioral variables, such curve fitting will not allow the researcher to draw firm conclusions about the processes that may or may not underlie the performance of a task.

We will now discuss more formally the issue of parameter variability and how it impacts on our ability to discriminate between alternative models. We wish to demonstrate that the analyses of RT data (or any other single dependent measure) in isolation from a model and from other dependent measures can sometimes lead to very different conclusions about the structure of the processes in a task and thus to emphasize the importance of a model-driven approach to data analyses.

Parameter Variability

The RT data collected for a particular experiment can be examined in a number of different ways, and the ones that we will focus on are the density, distribution, and hazard functions. Without the added burden of parameter variability, the hazard functions can be used to discriminate between different models (predicting different RT distributions), even if the density and distribution functions are very similar. With parameter variability, however, the hazard functions lose a great deal of their diagnosticity.

When we speak of parameter variability, we mean that the central process responsible for the execution of a task depends on a set of parameters that may be random variables. The way that the parameters vary may be systematic with respect to the experimental variables, or it may be random. On one trial, the rate at which a process executes might take on a value that remains fixed throughout that trial. However, during the next trial, the rate may be slightly different. Thus, an experimental manipulation need not influence the architecture of the process itself, but rather it might influence the way the process parameters vary from trial to trial. For instance, as a subject becomes more and more practiced at lexical decisions, the rate parameter of the process, say, a random walk, may systematically increase, leading to a decrease in processing time. Or the boundaries of the random walk might be adjusted as the subject becomes more familiar with the task.

Consider a memory search task. One model of the task might be that the addition of distractors to a list of items held in memory increases the number of comparisons between the memory list and a probe item. Each additional comparison is represented by a new subprocess required for each additional distractor item or, equivalently, the successive operations of some single neural module over the list items. (We make no distinction here between the serial or parallel nature of the comparison subprocesses.) An alternative model is one in which the addition of distractors increases the load on the central process responsible for recognizing the probe item. The addition of distractors might simply slow the retrieval process without requiring the addition of new subprocesses for each new distractor. For instance, the recognition process might slow with more distractors because they make the probe item less salient and, hence, more difficult to retrieve. As the number of distractors increases, RTs will increase just as if additional comparison processes had been added. This is a single-process model of memory search, where the addition of distractors causes a systematic change in a process parameter (salience), rather than an increase in the number of subprocesses (comparisons) required to complete the search. For both models, there may also be some nonsystematic variation of parameter values from trial to trial.

From within the context of a single experimental condition (say, a memory search procedure with five items), a question concerning the presence or absence of parameter variability or the number of processes involved in the task cannot be answered. It may be possible to find examples of several kinds of models that will fit the data arbitrarily well. Only across several experimental conditions would these issues become interesting and important. In changing the experimental situation, has the fundamental structure of the process changed, or has the value of a process parameter simply changed? The two memory search examples presented above provide an example of this type of distinction.

To illustrate this distinction more concretely, consider again these two memory search examples. In the multiple-subprocess model, each new comparison takes an additional amount of time to complete (an additional inter-completion time, if the comparisons are performed concurrently) and so is represented by the sum of a number of random variables. Each new distractor requires a

new random variable to represent its comparison time. Before new distractors are incorporated into the memory set, the total processing time is some random variable, say, T. After a distractor is added, then the total processing time is $T + T_1$, where T_1 represents the comparison time or intercompletion time for the new item. If the process is serial, then T_1 is the time to process the new distractor. If the process is parallel and the time for one comparison does not influence the other comparisons, T_1 is the time between the end of processing of the old items and the completion of the new item (the intercompletion time).

In the single-process model, the random variable T follows some distribution, $F(t|a)$, before the addition of new distractors. The nonvarying parameter a represents all of those critical aspects of the experimental situation that influence the processing time and, in particular, the salience of the probe item. The distribution F is determined by the architecture of the retrieval process. For example, if retrieval proceeds via a random walk or diffusion process, then F would represent the distribution of first passage times and the parameter a (in this case representing more than one parameter) would encompass the drift rate, starting point, drift variability, and so on (Ratcliff, 1978). After the inclusion of an additional distractor in the memory set, the total finishing time (still represented as T) would be distributed as $F(t|a')$, where the parameter a' now represents the new experimental situation and, in particular, the decreased salience of the probe item. In the diffusion model, decreased probe salience would result in a decrease in d' between old (familiar) and new (unfamiliar) items. Thus, the likelihood of selecting a small drift rate would be increased, and the RTs would therefore be longer overall.

The parameter a is a constant within a particular experimental situation and changes only when the experimental conditions change. It is also possible, however, to conceive of less systematic variations of the parameters. Momentary lapses of attention caused by unexpected noises outside the testing room add variability. Subject boredom or fatigue may lead to both systematic and nonsystematic deviations in the parameters across the course of an experimental session. The time of day that subjects are tested in multisession experiments can produce significant session effects. It is widely believed that subjects tested at the beginning of an academic semester or quarter can give very different data from those tested near the end (but see Langston, Ohnesorge, Kruley, & Haase, 1994). In the case where parameters change over blocks of trials, Burbeck and Luce (1982) have discussed a method whereby blocks with highly deviant parameters can be eliminated. However, if parameters are not constant within blocks of trials, there is no post hoc way to correct for their variability and, therefore, no way to observe the "pure" RT distribution.

Consider the best possible experimental procedure, where the best possible subject has performed under optimal conditions, the data has been carefully censored, and Burbeck and Luce's (1982) procedure for eliminating block of trials has been applied. Is parameter variability still a concern, or can we safely assume that we have eliminated all possible sources of variation and go on to examine the densities or hazard functions and draw conclusions about possible models? Even if all external sources of parameter variability have been removed, there may still remain one source of variability inherent in the perception of the stimulus. A unique stimulus never gives rise to exactly the same perceptual effect over the many times that it is presented. This idea is the foundation of psychophysics and signal detection theory. It also forms the cornerstone of most current models of information processing, perception, and memory. If we wish to presume that the performance of a task depends in some way on the percept of the stimulus, then the information upon which the process operates changes from trial to trial even if the stimulus remains the same. In this light, even the most carefully designed and executed experiment is susceptible to a not-insignificant degree of parameter variability. Over the course of an experiment, the experimenter collects many RTs arising from many different parameter values determined by the range of perceptual effects. Thus, it is quite likely that the data represent a statistical mixture from the possible RT distributions based on those perceptual effects.

The idea that RT data arises from some mixture of processes is not a new one. There have been several models that deal with effects in choice RT by assuming that on a particular trial the subject responds from one of several possible mental states and that different RT distributions are associated with each state (Falmagne, 1965; Falmagne, Cohen, & Dwivedi, 1975; Ollman, 1966; Yantis, Meyer, & Smith, 1991; Yellott, 1967, 1971). The observed distribution of RTs arises from a composite of all the different distributions generated by the separate mental states. These earlier models used a fixed number of states. In the two-state model, for example, the subject was assumed to have either some or no information about the stimulus presented. With some probability, p, on each trial, no information was gained and the subject then guesses; the guessing RTs follow some distribution. With probability $1-p$, some or all information was gained and the subject then performs the algorithm appropriate to the task; these RTs follow some other distribution.

Yantis et al. (1991), motivated as we are by concerns about parameter variability, have investigated the more complicated behaviors of multinomial mixtures, focusing specifically on the task of estimating the mixture probabilities in multinomial mixture models. In a multinomial mixture with N components, a parameter, α, follows a discrete probability distribution, and the observed RT distribution $F(t)$ is a weighted average of the processing-time distribution $G(t|\alpha)$:

$$F(t) = p_1 G(t|\alpha_1) + p_2 G(t|\alpha_2) + \ldots + p_N G(t|\alpha_N),$$

where the sum of the mixture probabilities $p_1 + p_2 + \ldots + p_N = 1$. Our interest is more general, however. What does the presence of mixtures imply about the RT distributions? We are also interested in the case where the

parameters of interest are continuous, such as any parameter would be that reflected stimulus intensity, for example. Models that incorporate such parameter variability produce continuous mixtures. Rather than a probability distribution defined over a finite number of possible mental states (e.g., p and $1-p$ for guessing and information-based responding, respectively), continuous parameter variability requires the specification of a continuous probability density function. The observed distribution $F(t)$ of the mixture is found by integrating the distribution $G(t|\alpha)$ of the processing times weighted by the parameter density $h(\alpha)$:

$$F(t) = \int_{\alpha \in A} G(t|\alpha) h(\alpha) d\alpha,$$

where A indicates the set of possible values the parameter α can take. For example, if a process gives rise to normally distributed finishing times (G), and the mean finishing time (α) of that process is exponentially distributed, then h represents the exponential density, A is the positive real line, and the observed RTs will be distributed as ex-Gaussian variables (F).

As we will discuss in the close of this article, one perspective on the problem of parameter variability is that it "reduces to absurdity." At what point may we stop worrying about parameter variability? For, indeed, the density function $h(\alpha)$ will itself require the specification of a number of parameters, which in turn may vary, and so on. Nonetheless, parameter variability may be modeled, and potential sources of variability are part and parcel of several areas of study in experimental psychology (e.g., Revelle, 1993). The evaluation of these sources and the way parameters are influenced by them is subject to the same constraints as the construction and evaluation of the models of the process under scrutiny.

We will return to this issue later on. For now, we will argue that tests that have been proposed to discriminate between different process architectures, such as serial and parallel processes, must also consider the case of a single process, possibly with variability in parameter values. In the discussion to follow, we will work through several theoretical problems. We will present the most specific model as a way to introduce some of the theoretical issues that will recur later in the paper. In this model, it is assumed that at least two subprocesses are operating in series and that the second subprocess gives rise to exponentially distributed processing times (Ashby & Townsend, 1980). We will demonstrate that, even in this most specific case, single-subprocess models with parameter variability produce results that are indistinguishable from the inserted exponential subprocess model. We will then move to a discussion of the logic of additive factors, in which it is assumed that at least two subprocesses are operating in series. We will then consider two established models and demonstrate the existence of competing models that cannot be distinguished from their alternatives on the basis of RT data alone.

CASE 1
Testing for the Presence of an Exponential Serially Inserted Subprocess

Ashby and Townsend (1980) have worked extensively with one specific, simple multiple-subprocess model of information processing. Consider an experimental design in which an independent variable of interest (factor) takes on some number of levels; for example, in a memory search paradigm, the factor under study might be the memory set size. At level $k-1$ of this factor ($k-1$ items held in memory), the RTs follow some distribution, F_{k-1}. Increasing the level of this factor to k (by adding another item to the memory set) slows the RTs, now distributed as F_k. If this increase to level k caused the insertion of an additional subprocess (for example, an additional comparison between the probe and the memory set), and the duration of this subprocess is exponentially distributed and independent from the RTs at level $k-1$, Ashby and Townsend showed that the RT density at level k (f_k) is proportional to the difference between the distribution functions at levels $k-1$ and k. Or

$$f_k = v_k (F_{k-1} - F_k). \tag{1}$$

The constant of proportionality v_k is the rate of the inserted exponential subprocess. This result provides an empirical test of the exponential stage model. After estimating f_k, F_{k-1}, and F_k from the data, the plot of f_k versus $F_{k-1} - F_k$ can be examined for linearity; the slope of the resulting line gives an estimate of the inserted exponential processing rate. Or, equivalently, the ratio $f_k/(F_{k-1} - F_k)$ can be plotted as a function of time; if the inserted exponential subprocess (IES) model is true, the slope of this function will be approximately zero, and the intercept will give an estimate of the exponential processing rate. This test is particularly advantageous in that no assumptions are required about the shape of the distribution F_{k-1}. If an IES model of the task is correct, the proportional relationship will hold regardless of the original distribution F_{k-1}.

Ashby and Townsend (1980) noted, however, that "there may be models, not formally equivalent to [the IES model], capable of predicting functions close enough to being flat and linear that an empirical application of the [IES test] could not determine that they were not" (p. 101). In an attempt to determine the statistical power of the test, they conducted several simulation studies, in which an additional subprocess was distributed as either an exponential or some other positive random variable. From the slopes of the regression lines through $f_k / (F_{k-1} - F_k)$ observed when the additional subprocess was exponential versus when it was not, they suggested that a good criterion for rejecting the IES model might be when the absolute value of the slope was greater than 1×10^{-4}/millisecond. They estimated the RT density and distribution functions from a published set of rapid memory search data (Townsend & Roos, 1973). For 3 subjects at five processing loads each, they found that the data passed

the IES test in most cases. For these data, only 3 of the 12 regression equations had slopes with absolute values that exceeded 1×10^{-4}/millisecond. The results of this analysis provide support for a multiple-subprocess model of rapid memory search, in which the addition of a distractor item to the search set induces an additional, exponentially distributed comparison process. However, as Ashby and Townsend presaged, other single-process models can be formulated that pass the IES test. For example, Ratcliff (1988) showed that a diffusion model produces RT distributions that meet the IES criterion using parameter values derived from fits to empirical data.

We have fit several other models to RT distributions based on the IES model. First, we generated the curves for an IES model observed over three "experimental" conditions. In the condition producing the fastest RTs, the process was composed of two independent serial stages, each finishing with times that were exponentially distributed with equal means. The total processing time for this condition was then the sum of two independent and identically distributed exponential variables. In the second condition, a new exponential stage was added after the first two stages. This new subprocess was independent from the original (multiple) process, but identically distributed to the original exponential stages, yielding total processing times that were the sum of three exponential random variables. The RTs from the last condition resulted from inserting still another independent, identical exponential stage after the first three, giving total processing times that were the sum of four exponential variables. Because the second and third conditions resulted from inserting new, independent exponential subprocesses, it must be the case that (1) the RT density in the second condition is proportional to the difference between the cumulative distribution functions for the first and second conditions, satisfying Equation 1, and (2) the RT density in the third condition is proportional to the difference between the cumulative distribution functions for the second and third conditions, also satisfying Equation 1.

After generating the densities for the IES model, we then fit the densities predicted by two other single-process models to the IES densities. The first of these, suggested by Hohle (1965), was a single exponentially distributed "decision" stage that slowed across conditions, with an additional, normally distributed time component absorbing the times for perception, response execution, and any other stages not influenced by the experimental manipulation. The second of these, discussed earlier, arises from a diffusion process with a single absorbing boundary, with a drift rate that decreased across conditions. The fits of these models to the IES RT densities are shown in Figure 3. The top panel shows the fits of the single exponential stage model (producing ex-Gaussian densities), and the bottom panel shows the fits of the single-boundary diffusion model (producing inverse normal densities).

We subjected the single-process models' RTs to the IES test. The ratios of the densities to the differences between the distribution functions are shown in Figure 4, along with the ratios for the IES model. Within the usual observed range of RTs, the single-process models pass the IES test, even though no additional exponential stages were added.

These two single-process models used a systematic, nonrandom shift in the process parameters from one experimental condition to another to mimic the behavior of a multiple-process model. Although they pass the IES test, demonstrating statistical mimicking at the level of the density and distribution functions, it is possible that other aspects of the data can be used to distinguish between them, such as the hazard function. The hazard functions for each model in each condition are presented in Figure 5, along with the hazard functions of the IES model. There are clear differences between the hazard functions for the single- and multiple-process models, so the hazard functions might be used to discriminate between them.

However, if parameters are variable from trial to trial, it may no longer be possible to use the hazard functions

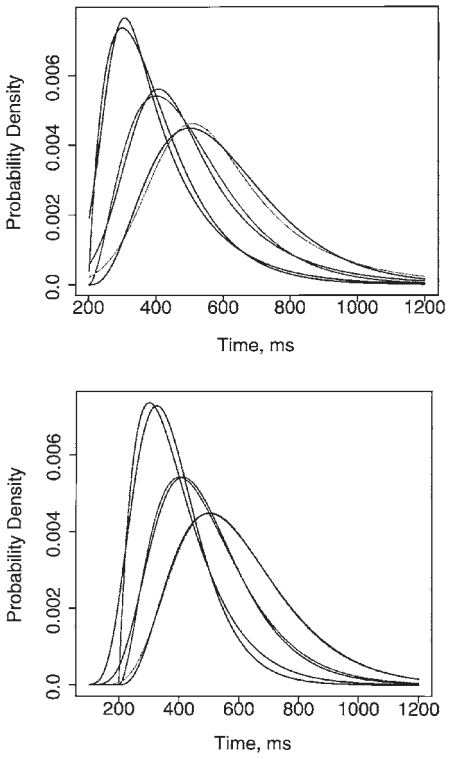

Figure 3. Mimicking the IES model with single-process models. The top panel shows three ex-Gaussian densities fitted to three gamma densities. The gamma densities have a rate of .01 and a base time component of 200, and shape parameters of 2, 3, and 4. The normal components of the ex-Gaussians have means of 247, 322, and 402 msec, and standard deviations of 52, 78, and 104 msec, and the exponential components have means of 148, 182, and 203 msec, for shape parameters 2, 3, and 4, respectively. The bottom panel shows three inverse normals fitted to the same three gamma densities. These inverse normals have means of 380, 490, and 597 msec, and λ parameters of (see Luce, 1986, p. 509) 3,610, 4,194 and 5,244 msec for shape parameters 2, 3, and 4, respectively.

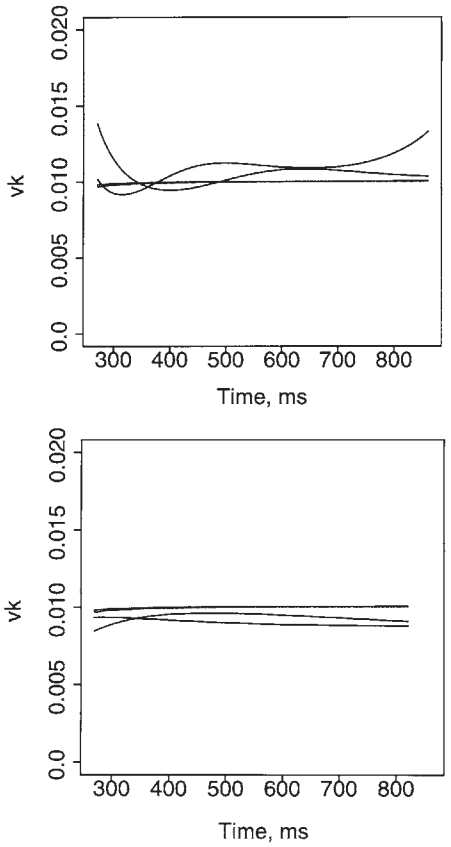

Figure 4. The results of the IES test for the ex-Gaussian (top panel) and inverse normal (bottom panel) densities shown in Figure 1. Each plot shows the value of the IES ratio $f_k(t)/[F_{k-1}(t)-F_k(t)]$ over time. The absolutely flat lines in the figures are the ratios for the IES (gamma) model. The not-quite-flat lines are the ratios for the mimicking models.

in this way. To provide a concrete example of the problem of parameter variability, we present the following scenario: Suppose that the time for completion of some central process of interest is determined by the length of time that a stimulus remains visible on a computer display. The central process proceeds in exactly the same way regardless of the duration of the display, producing for a given duration a normally distributed finishing time of some mean and variance. If the duration is very short, for example, the mean finishing time might be very long, producing very slow, highly variable RTs.

A researcher is testing an alternative model that assumes that when a stimulus is presented for a short period of time, the subject performs an additional "rechecking" procedure, reexamining the evidence that led to the central process outcome to make sure that a selected response is appropriate. When the display is presented for longer durations, the rechecking procedure is unnecessary, and so the rechecking subprocess is not executed. He has two conditions in his experiment, a short and a long stimulus duration. However, there is a bug in the experiment program and the duration of the stimulus is actually very short, short, long, or very long, with the durations that he desires lying somewhere between these four actual times. Because the decision process is producing normally distributed RTs based on the display duration, for this faulty program, the collected RTs arise from one of four normally distributed variables, each with a different mean and variance corresponding to the different durations. The duration manipulation, unbeknownst to our researcher, makes some display durations more likely than others, but it does not shift the duration from a constant long to a constant short display as he believes.[2]

What occurs on each trial is then a sampling procedure, where a duration is selected at random from one of the possible four. This results in parameter variability from trial to trial. In this kind of parameter variability, the distribution of the parameter is discrete; at any time, the mean and variance of the process take on a set of values with some probability, which is equal to the probability of a particular display duration. The discrete probability distribution of durations represents the mixture proportions. When the experimenter increases the display duration, he changes those mixing proportions and so influences the parameter distribution, not the structure of the decision process itself.

Consider the normal distributions presented in the top panel of Figure 6. From fastest to slowest, these distri-

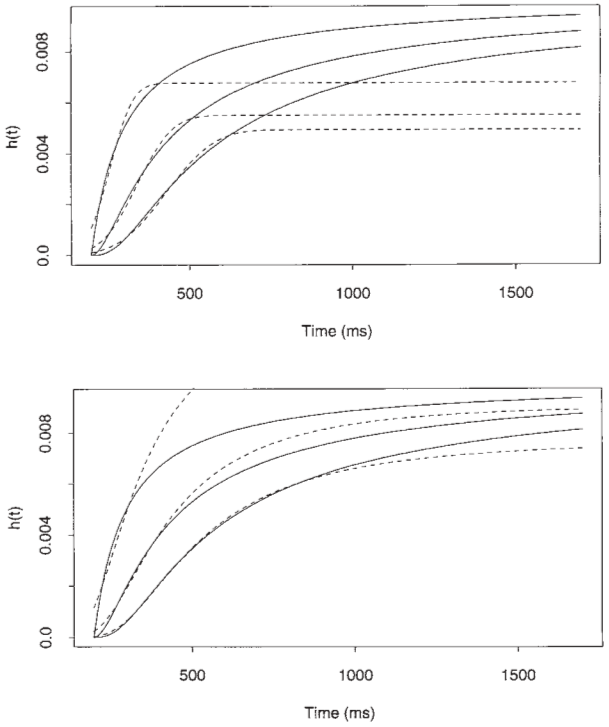

Figure 5. The hazard functions of the models shown in Figure 3. The top panel shows the hazard functions of the ex-Gaussian densities (dashed lines) superimposed on the gamma hazard functions (solid lines). The bottom panel shows the hazard functions of the inverse-normal densities (dashed lines) superimposed on the gamma hazard functions (solid lines).

butions have means of 400, 500, 600, and 700 msec, and standard deviations of 50, 150, 200, and 250 msec, respectively. The longest displays result in the fastest RTs, and the shortest displays result in the slowest RTs. When the experimenter presents "long" displays to the subject, the subject observes the slowest displays with a probability of .43, the next slowest with a probability of .20, faster displays with a probability of .01, and the fastest with a probability of .36. When the experimenter "decreases" the display duration, no additional rechecking stage is performed by the subject (i.e., there is no effect on the central process, which still finishes with the same normally distributed times as before), but the likelihood that a subject will be presented with any of the four actual durations is changed. Suppose now that, in this case, the subject receives slower, slow, fast, and faster displays with probabilities of .07, .22, .21, and .50, respectively. The data collected from this subject for "long" and "brief" displays are shown in the bottom panel of Figure 6 (which depicts the two mixture densities resulting from a simulation of the two conditions, with 1,000 observations at each level). All observations faster than 200 msec and slower than 1,200 msec (less than 1% of the total data points) were eliminated before using a Gaussian kernel technique (Parzen, 1962) to estimate the densities. The RT distribution functions were estimated with the cumulative relative frequency distributions of the observations from both conditions.

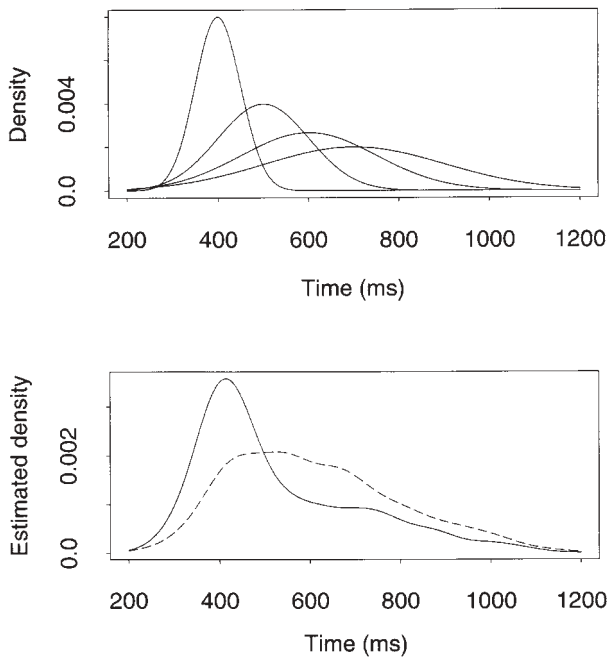

Figure 6. Top panel: The four component normal densities produced by each of the four actual display durations. The leftmost density with a mean of 400 msec is of processing times for the longest displays; decreasing display durations shift the density to the right and increase the variance. Bottom panel: The two estimated densities observed under the two display-duration conditions, resulting from two different mixtures of the densities shown in the top panel. The solid line indicates the observed RT density for the long display duration, and the dotted line the observed RT density for the brief display duration.

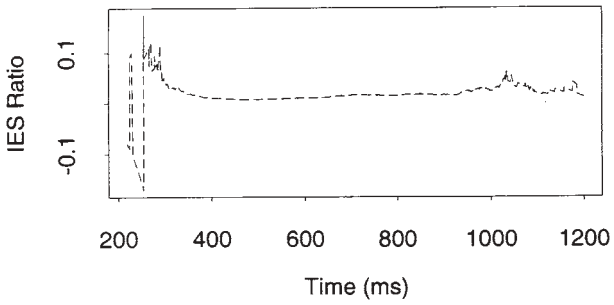

Figure 7. The IES ratio $f_k(t)/[F_{k-1}(t) - F_k(t)]$ produced by the mixture densities observed in the "long" $[F_{k-1}(t)]$ and "brief" $[F_k(t)]$ conditions.

To test his hypothesis of a rechecking stage, he plotted the ratio of the short duration density to the difference between the long- and the short-duration distribution as a function of time. This plot is shown as the dashed line in Figure 7. Although there is some variability in the plot in the extremes of the range, the ratio appears quite flat. Indeed, his regression analysis of this function yielded a slope of 1.7×10^{-5}—well within Ashby and Townsend's (1980) suggested limits of $\pm 1 \times 10^{-4}$. He thus concluded that shortening the duration of the display induced the subject to perform an additional rechecking procedure and that this rechecking time is exponentially distributed with a rate approximately equal to .006. To verify this estimate, he plotted the linear relationship between the short display density and the difference between the distributions, and another regression analysis gave an estimated slope of .008. Observing that the mean RT was slowed 84 msec by "decreasing" the display duration, he compared these values with $1/84 = .012$ and concluded that they were sufficiently consistent, putting the estimated processing rate somewhere around .01 or the duration of the rechecking procedure at around 100 msec.

Let us now compare his results with the results of another simulation in which an IES is present in the short display duration. In this simulation, the 84-msec mean increase arose from the addition of an exponential subprocess with a rate of 1/84. So, the RT distribution collected for the short displays was the result of adding an exponential variate to the mixture distribution observed at the long display (a new simulation of the long-display mixture was performed). The density of the times produced by the original mixture plus an IES was estimated as before, and the densities for both the mixture and the IES models at the short display duration are presented together in Figure 8. Although the leading edges and the tails of the two densities are quite similar, there is considerable disparity around the peak. The IES model density peaks more sharply than does the mixture model density. Unfortunately, this post hoc observation provides no way to distinguish between the IES and other

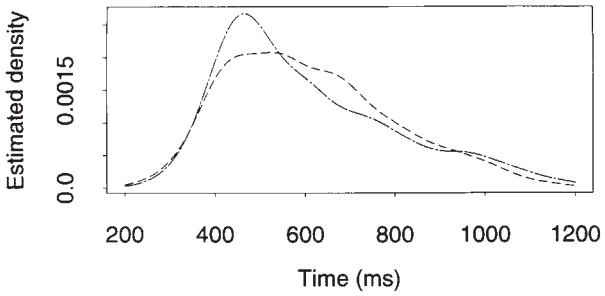

Figure 8. The estimated mixture density produced by the brief display duration (dashed line; the mixture model) superimposed on the estimation of the density produced by the long display duration plus an additional exponential stage (broken line; the IES model).

models in the course of a real experiment; that is the purpose of the Ashby and Townsend (1980) test.[3]

The same regression analyses were performed as on the "experimental" data. The IES ratio was calculated and is presented in the top panel of Figure 9. Regression analysis gave an estimated slope of 4.2×10^{-7}, well within Ashby and Townsend's (1980) prescribed limits and smaller than that observed for the mixture model. The exponential rate estimated from the intercept of this analysis was .017; the slope for the regression analysis performed on the density and the distribution differences was .011. The actual exponential rate was .012. The IES test produces no more consistent estimates for the exponential processing rate in the case where an IES was actually present than when it was not, and both models pass the IES test. The IES ratio for both models are presented together in the bottom panel of Figure 9. Clearly, the mixture model ratio is flat and very close to the ratio produced by the IES model. The tails of the mixture model are noisier than are those of the IES model; however, over approximately 95% of the range of the data, the two models are indistinguishable using this test.

Hazard Function Analyses

The preceding discussion has focused on the shape of the RT densities. As we discussed earlier, because empirical RT distributions have no remarkable characteristics that can aid in the discrimination between different classes of positive random variables to which RTs might be assigned, Luce and others have suggested that the RT hazard functions might give more diagnostic information about the class of random variables to which RTs belong (Bloxom,1984; Burbeck & Luce, 1982; Luce, 1986). The hazard function $h(t)$ is the ratio of the density function $f(t)$ to one minus the distribution function $F(t)$:

$$h(t) = \frac{f(t)}{1-F(t)}$$

We will use the previous IES example to illustrate some of the difficulties that arise in the estimation and subsequent use of the hazard function. Recall that the four normal variables arose from a probability distribution over four possible display durations. When the display duration was shortened in the faulty experiment program, the probability distribution changed to include a larger proportion of shorter displays. The data from the experiment passed Ashby and Townsend's (1980) test for the inclusion of an exponential serially inserted stage, the IES model. We now wish to determine if the hazard functions estimated from the data can be used to discriminate between the mixture distribution and the IES model. This is a reasonable approach to take: each normal distribution composing the mixture has a monotonic increasing hazard function, whereas the exponential components of the IES model have constant hazard functions. We might therefore expect that the shapes of the hazard functions for these two alternatives might be very different.

The hazard functions for the short-duration condition, as collected in the "experiment" (as predicted by the mixture) and as simulated by the IES model, are presented in Figure 10. There are differences between the two curves, especially in their tails. The IES hazard function appears to increase throughout the range of the data, whereas the mixture hazard function appears to increase and then decrease. Unfortunately, this difference in monotonicity is generated from 1% of the data: 6 observations from the "experiment" and 14 observations from the simulation. Thus, no conclusions can be drawn from these hazard functions; the IES model still cannot be distinguished from the mixture. The hazard-function estimate for the mixture data is, in fact, in error: the actual hazard function for this particular normal mixture is monotonic increasing (see Figure 11). In Figure 11, the

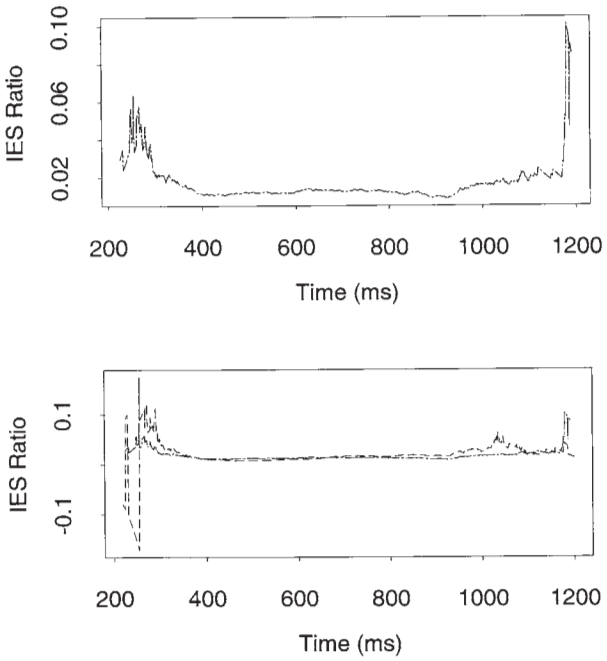

Figure 9. Top panel: The IES ratio produced by the addition of a serial exponential subprocess to the mixture process of the long-display condition: the IES model. Bottom panel: The mixture model (Figure 7) and IES model (top panel) ratios superimposed.

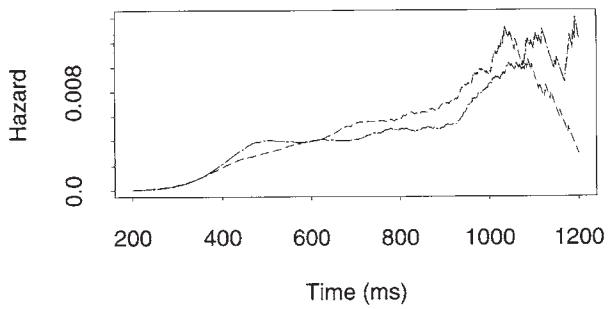

Figure 10. The estimated hazard functions of the densities shown in Figure 8. The mixture model (dashed line) shows a relatively constant increase over time, whereas the IES model (broken line) shows an early rise to asymptote followed by a later increase.

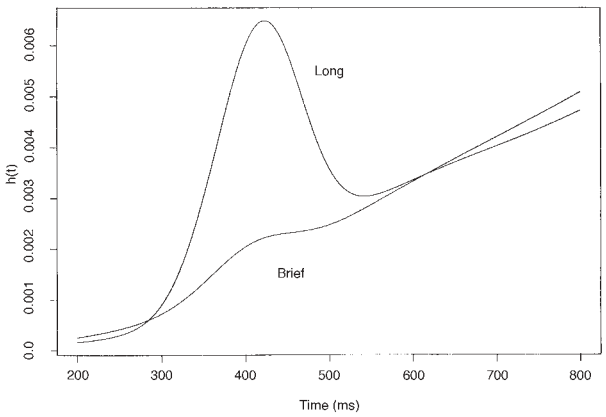

Figure 11. The actual hazard functions of the two normal mixtures presented in Figure 6's bottom panel. The "long" hazard function is formed from many fast RTs, whereas the "brief" hazard function contains many more slow RTs.

theoretical hazard functions for the two mixtures are presented. The hazard function for the short-duration condition is nondecreasing, whereas the hazard function for the long-duration condition is increasing then decreasing. This is typical of RT data (Ashby, Tein, & Balakrishnan, 1993; Balakrishnan & Ashby, 1992; Heath & Wilcox, 1990). More difficult conditions often show nondecreasing hazard functions across the range of data; however, as the stimulus becomes more intense or more easily detectable, the peak of the hazard function moves up and to the left (Burbeck & Luce, 1982; Luce, 1986).

Burbeck and Luce (1982) used the generally increasing then decreasing shape of the hazard function to argue for a two-component process in auditory detection. The rapid peak of the function was tied to an early "change" detector, and the later asymptote to a more slowly operating "level" detector. The simultaneous operation of these two detectors, they argued, gave rise to the increasing then decreasing hazard. They compared a two-detector model of auditory detection with a random-walk model (predicting an inverse normal distribution) and Grice's (1968) accumulator model. All three of these models predict that the hazard functions should either increase monotonically to asymptote or peak and then decrease to asymptote. Comparing the hazard function predictions of the three models with their observed empirical hazard functions, they found in favor of the two-detector model. As they mentioned, and in light of the previous demonstration, an alternative explanation is that their data was formed by a mixture of processes arising from parameter variability.

Because of the mixture problem, Burbeck and Luce (1982) gave serious consideration to the question of parameter variability in the course of their investigation of the two-detector model, especially as it occurs between blocks of trials or sessions in a multisession experiment. They presented a method for discarding blocks of trials in which the parameter values have clearly drifted beyond those of other blocks. They acknowledged, however, that correcting for parameter variability (mixture distributions) arising within a block of trials is impossible and that the peaking and then decreasing to asymptote behavior of the hazard functions may also arise from a mixture of distributions within blocks.

The nonmonotonic characteristics of mixture hazard functions was also discussed at length by Barlow and Proschan (1975). An example of the way that hazard function analyses can result in ambiguous conclusions was provided by Proschan (1963). He examined the pooled failure times for air-conditioning systems in 13 Boeing aircraft. He assumed that the failure times were exponentially distributed but noted that the hazard function of the pooled data decreased. Because the exponential hazard is constant, he commented briefly that, although a mixture of exponential rates could have led to this type of hazard function, perhaps the original assumption of exponentially distributed failure times might be questioned on the basis of this finding. Dahiya and Gurland (1972) subsequently demonstrated that a gamma distribution with the shape parameter (number of stages) less than one could be

well fit to the data, suggesting that perhaps the failure times were gamma-distributed. To reconcile these two findings, Gleser (1989) noted that a gamma distribution with a shape parameter less than one can be represented as a mixture of exponentials with different rates. He pointed out that the empirical hazard function gives no basis for ruling out the original exponential assumption, since mixtures of exponentials can give exactly the kind of hazard functions that Proschan observed. Thus, the shape of the hazard function is ambiguous with regard to the original distribution when a mixture is possible.

Hazard-function analyses are becoming more common in areas of psychology apart from RT research. For example, Lewinsohn, Zeiss, and Duncan (1989) tallied the frequency of episodes of unipolar depression for several clinical groups. All groups showed an increasing likelihood of experiencing an episode over time, as reflected in an increasing hazard function; however, they observed that men with only one prior depressive episode eventually declined in vulnerability to a second episode (i.e., showed an increasing then decreasing hazard function as time progressed). They then went on to discuss the implications of such a finding, in terms of clinical strategies and theoretical issues. Levinthal and Fichman (1988), while investigating the likelihood of maintaining an auditor–client relationship, called the initial rise of the increasing then decreasing hazard function a "honeymoon period in the first few years of attachment" (p. 355). In short, the increasing then decreasing hazard is ubiquitous; it appears not only in RT data but in almost all sets of data collapsed across subjects or observations collected over significant periods of time. This often-observed increasing then decreasing shape is not necessarily indicative of the underlying process if there is any possibility that the data results from a mixture, whether that is a "honeymoon" period, a sudden remission from clinical depression, simultaneous operation of fast and slow detectors, or something else. It may indicate that the data set is composed of a mixture of two or more distributions or indicate the presence of a parameter varying across trials.

Balakrishnan and Ashby (1992, p. 82) state that nonmonotonic "empirical estimates of [the hazard function] unequivocally rule out the traditional candidates for the RT distribution, including the gamma, log-normal, and ex-Gaussian." Each of these "traditional candidates" has a nondecreasing hazard function. By implication then, can we rule out all process models that predict gamma, log-normal, or ex-Gaussian finishing times? We may if the data could not possibly have arisen from a mixture process. However, if the data result from observations from, say, a process producing gamma-distributed finishing times under several different rates, the resulting hazard function may indeed be nonmonotonic.

To clearly illustrate how mixtures produce nonmonotonic hazard functions, the top panel of Figure 12 shows the hazard functions of four gamma distributions (solid lines), each of a different rate but with the same shape parameter, which are monotonic increasing over the range presented. Superimposed on these functions is a hazard function of a mixture of these four distributions (dotted line). The bottom panel of Figure 12 shows another example, using the hazard functions of normal distributions. Notice that the mixture hazard function is still increasing and then decreasing, even though the component hazard functions are positively accelerated. The increasing then decreasing pattern is formed from the way that the hazard of the mixture "tracks" the behavior of the hazard functions of the distributions from which it is composed. Early in the process, the mixture hazard function follows the behavior of the fastest distribution; later in the process, it follows the slowest. This demonstrates that the increasing then decreasing hazard does not in fact rule out models predicting either the gamma or the normal classes of distributions of processing times, unless one is willing to suppose that all possible sources of parameter variability have been eliminated from one's paradigm. As we argued earlier, this is highly unlikely and indeed inconsistent with the use of the theoretical tools of signal detection theory or the notion of perceptual variability.

Summary

This section has addressed a particular model of information processing, the IES model. This model makes no presumptions about the initial structure of a process; however, upon the manipulation of some variable of interest, it assumes that an additional subprocess is added, leaving the original process unchanged, and that the finishing time of this new subprocess is exponentially distributed. The new finishing time is then distributed like the old finishing time plus an exponential random variable. This model makes a very specific prediction concerning the relationship between the pre- and postmanipulation RT distributions, and this prediction provides a strong test for the IES model. Because this model assumes very definite structural changes across conditions, and because the relationship between the distributions must therefore be very constrained, it may at first seem unlikely that there are models other than the IES model that would satisfy the distributional conditions. Also, the IES test is bidirectional: the finding of a flat IES ratio is logically equivalent to the insertion of a serial exponential stage. Thus, it is true that no other type of model could produce a truly flat ratio. Nonetheless, there are several types of distributions that can be statistically indistinguishable from those predicted by the IES model.

In some instances, when the parameters of the process are constant, the RT hazard functions can be used to discriminate between the different models. However, when parameter variability is introduced, producing mixture distributions of RTs, this is no longer the case. We showed that mixture distributions can also produce the flat IES ratio and, furthermore, that the hazard functions are no longer very useful for testing between the models.

The IES model makes the assumption of exponentially distributed processing times primarily for mathematical tractability and not from specific process considera-

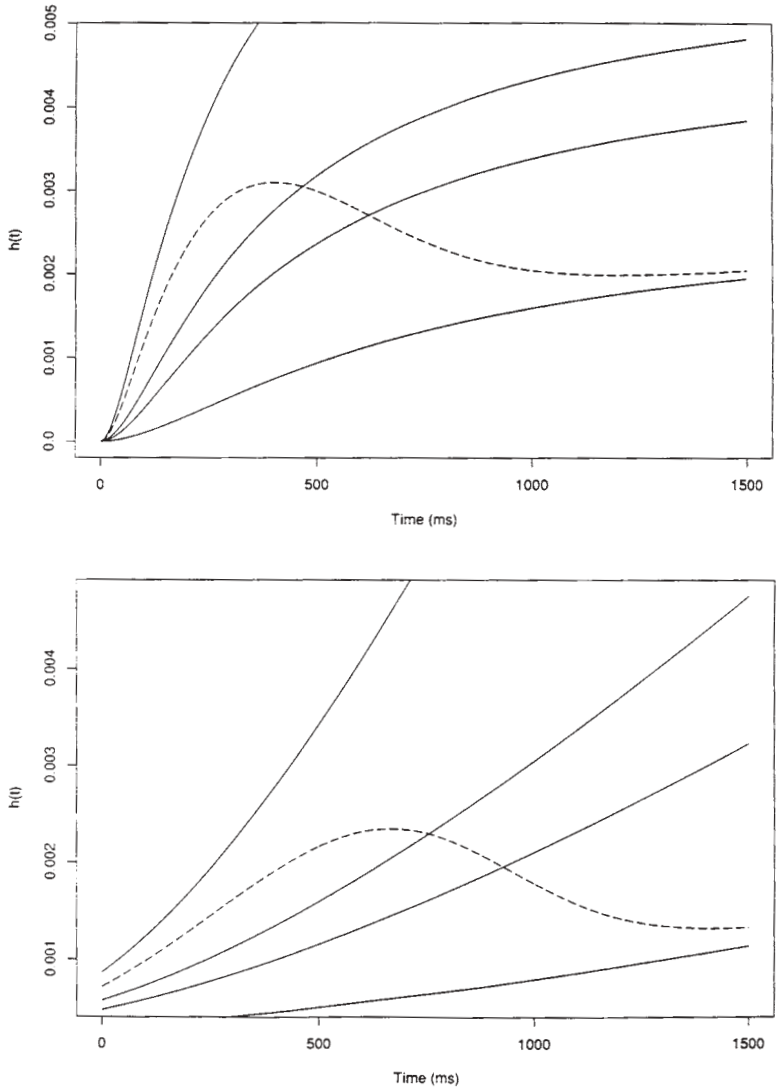

Figure 12. Top panel: The hazard function of a mixture of four gamma densities. The component hazard functions are shown as solid lines, and the mixture hazard function is shown as a dotted line. Bottom panel: The hazard function of a mixture (dotted line) of four normal densities (solid lines).

tions. This is by no means a shortcoming: tractability is a significant concern, and even if it was not, there are a number of other grounds on which such an assumption could be justified. However, we have shown that there exist other structures that do not include exponential subprocesses that behave in the same way as the IES model. To a large extent, the reason for this is that we are only considering RTs. The IES model and most of the alternative models considered make only RT predictions, so we cannot compare the predictions of the alternative models and the IES model for other behavioral variables.

Ashby and Townsend (1980) also have considered a more general model, in which two critical stages of processing operate during the performance of a task. This is the classic serial stage model, around which Sternberg (1969) developed the additive factors logic. The assumption is that manipulating the levels of certain experimental factors serves to extend the processing time for separate stages, and, therefore, the arrangement of those stages can be determined from the resulting pattern of mean RTs. Ashby and Townsend extended the additive-factors logic to the entire RT distribution and suggested a mathematical test for the presence of serial stages. Roberts and Sternberg (1994) recently developed an empirical version of this test and applied it to several sets of data. We will now turn to a discussion of additive factors and the Ashby and Townsend test to investigate the problem of statistical mimicking for this larger class of models.

CASE 2
Additive-Factors Method and Its Extension to RT Distributions

Sternberg (1969) presented what has come to be known as the additive-factors approach to testing cognitive models. Given some presumed number of subprocesses involved in a task, and some assumptions about which experimental manipulations affect those subprocesses, Sternberg proposed that the interactions in mean RT data could provide information about how the processes were arranged. In particular, if two factors have additive effects on mean RT, if they do not interact, and if we have reason to believe that the two factors are influencing two different stages of processing, we might assume that the two stages were arranged serially: one stage completes before the other begins.

In an illustration of this logic, Sternberg (1967) performed an experiment in which he manipulated two factors in a memory search task. Each factor was assumed to affect one and only one stage of processing—that is, the factors had selective influence. The first factor was the visual integrity of the probe item (e.g., masked or unmasked), which was assumed to affect only the stage of stimulus encoding and leave the subsequent memory search process unchanged. The second factor was the number of elements in the memory search set, which was assumed to influence the comparison stage between the probe item and memory but leave the stimulus encoding stage unchanged. He used two levels of probe integrity (high and low), combined with several memory loads. If the subprocesses of stimulus encoding and memory search are arranged serially, so that the search process cannot begin until the probe is encoded (and if the durations of these two stages are stochastically independent), then the mean RTs should not have interacted for integrity and load. That is, the means should have been additive: the increase in mean RT under low integrity, high load should have equalled the sum of the mean RT increases when only integrity was decreased and when only load was increased. Indeed, Sternberg observed an additive pattern of mean RTs over three memory loads, thus providing evidence consistent with the hypothesis that the processes of encoding and search were arranged serially.

The use of additive mean RTs as support for a serial arrangement of subprocesses depends heavily on the assumption of selective influence of the factors manipulated. If the manipulations of interest influence a common subprocess, or if a third factor exists that influences both subprocesses simultaneously, additivity of the mean RTs may not hold even if serial subprocesses are present (see Townsend & Ashby, 1983, for a discussion). If a third subprocess operates concurrently with the two processes of interest, even more complicated patterns of mean RT can be observed (e.g., Schweickert, 1978, 1980, 1983). Furthermore, Townsend and Ashby (1983) have outlined classes of parallel models that also produce additivity. Townsend and Thomas (1994) have also presented many implications of the failure of selective influence on different processing structures. In sum, the additive-factors method applied to mean RTs alone does not provide unambiguous information about the arrangement of subprocesses in a task.

So far, we have discussed statistical mimicking at the level of the RT distributions. Thus, the comparisons between means, medians, standard deviations, and so on all followed as a function of the goodness of match between the distributions predicted by the different models. Statistical mimicking at the level of the means is much easier to accomplish. Indeed, it usually is not necessary to even specify the distributions from which the means are measured (see, e.g., Townsend & Ashby, 1983; see also Appendix A). Under quite general conditions, and in the absence of other variables, there exists a large number of reasonable alternative models that will produce a given pattern of mean RT data. Therefore, the additivity or lack thereof in a particular set of mean RT data does not rule out the possibility that some other kind of model that does not have serial subprocesses was acting to produce the RTs.

Ashby and Townsend (1980) presented a more rigorous test for the presence of serial subprocesses by extending the additive-factors logic to the level of the RT distribution functions. Suppose there are two critical subprocesses of a task (a and b), such as the stimulus encoding and memory search processes discussed above. Suppose also that we can find two experimental factors (A and B) that have selective influence on subprocesses a and b, respectively, such as the visual integrity and memory load manipulated in Sternberg's (1967) experiment. If the subprocesses a and b are arranged serially, then the total processing time (not including the time for processes that remain unchanged throughout the experiment) can be expressed as the sum of two random variables, $T_{ai} + T_{bj}$, representing the separate durations for subprocesses a and b, respectively, when factor A is at level i and factor B is at level j. Notice that this notation restricts the influence of factor A to T_a and factor B to T_b: selective influence.

If factors A and B each have two levels, then a factorial experimental design gives four random variables for the processing times in each condition: T_{11}, T_{12}, T_{21}, and T_{22}, where T_{ij} indicates the processing time when factor A is at level i and factor B is at level j. If a and b are arranged serially, each T_{ij} can be broken down as a sum of its component subprocessing times (for example, $T_{11} = T_{a1} + T_{b1}$). Examining this decomposition for each T_{ij} shows, for the 2×2 factorial design, it must be the case that $T_{11} + T_{22} = T_{12} + T_{21}$. Therefore, as Ashby and Townsend (1980) noted, the distributions of two random variables defined by the sum of the times observed in the 11 and 22 conditions and the sum of the times observed in the 12 and 21 conditions should be equal if selective influence and seriality hold. That is,

$$F_{T_{11} + T_{22}}(t) = F_{T_{12} + T_{21}}(t). \tag{2}$$

Roberts and Sternberg (1994) developed a straightforward empirical application of this condition. By adding each RT collected in the 11 condition to each RT

collected in the 22 condition, and similarly for the 12 and 21 conditions, samples drawn from the distributions of $T_{11} + T_{22}$ and $T_{12} + T_{21}$ are formed. The distribution functions of these variables can then be estimated, and the truth or falsity of Equation 2 can be checked. Roberts and Sternberg called this procedure the *summation* test.

Roberts and Sternberg (1994) also presented a distributional test for two-component mixtures: in particular, the *alternate-pathways* model. In the alternate-pathways model, subprocesses *a* and *b* are arranged in such a way that, on some proportion of the trials, process *a* is performed and, on the remainder, *b* is performed. Unlike the mixture models we have discussed to this point, the mixing proportions (probabilities of performing *a* or *b*) do not change from one condition to the next. The distribution function $F_{ij}(t)$, the distribution of the RTs in condition *ij*, must equal $pF_{ai}(t) + (1-p)F_{bj}(t)$, where p is the probability of executing process *a*, or taking path *a*, and $F_{ai}(t)$ and $F_{bj}(t)$ are the processing-time distributions for each path under condition *ij*. The value of p does not change with the level of either *A* or *B*. Instead, as for the serial stage model, factors *A* and *B* selectively influence *a* and *b* such that if *A* is increased, process *a* slows, and if *B* is increased, process *b* slows. For this model, the distribution functions for the separate conditions must satisfy

$$\frac{1}{2}[F_{11}(t) + F_{22}(t)] = \frac{1}{2}[F_{12}(t) + F_{21}(t)]. \qquad (3)$$

The distribution functions for the conditions can be estimated and averaged in pairs to form the left and right hand sides of Equation 3. The truth or falsity of this mixture test can then be determined.

Roberts and Sternberg (1994) noted that mathematically, no set of distributions could satisfy both the mixture and the summation tests. Thus, a set of data that passes the summation test logically must fail the mixture test, and vice versa. They performed many analyses on both the distributions and the distribution statistics from five data sets and found that the summation test held throughout much of the data, whereas the mixture test failed. They also examined the distributional predictions of the cascade model (Ashby, 1982; McClelland, 1979). Because the data did not pass the mixture test or fulfill the predictions of the cascade model, their analyses provide evidence consistent with serial stage models of the various tasks that they investigated.

The most attractive feature of the summation and the mixture tests is that they are not tied to any particular specification of the distributions of the random variables involved. Just as the IES test is, these tests are distribution-free. If the RTs were generated by a serial stage or alternate-pathways model, either Equation 2 or Equation 3 must hold regardless of the shape of the functions $F_{ij}(t)$ or the value of the pathway probability p. Theoretically, one should be able to apply these tests to the data and be able to draw conclusions about basic cognitive architecture (such as the serial arrangement of processes) without an elaborate model of the subprocesses involved. However, consider the summation test from a different perspective. Notice that a model that will pass the summation test is one where the sums of T_{11} and T_{22} and of T_{12} and T_{21} are equal in distribution. It is not necessary to assume that each of those four random variables are themselves composed of sums of identical random variables.

Suppose, for example, that all T_{ij}s are independent and normally distributed. T_{11} has a mean of 350 msec and variance of 1,000 msec^2, T_{12} has a mean of 450 msec and variance of 2,000 msec^2, T_{21} has a mean of 550 msec and variance of 3,000 msec^2, and T_{22} has a mean of 650 msec and variance of 4,000 msec^2. Then, because sums of normal random variables are normally distributed, $T_{11} + T_{22}$ and $T_{12} + T_{21}$ are equal in distribution: both sums are normally distributed with a mean of 1,000 msec and variance of 5,000 msec^2. Therefore, these distributions will demonstrate mean and variance additivity and pass the summation test. Each normally distributed variable can of course be broken down into sums of two other normal variables in such a way that the $T_{ij} = T_{ai} + T_{bj}$ representation is sensible. For example, if T_{a1} had a mean of 200 msec and variance of 500 msec^2, T_{a2} had a mean of 400 msec and variance of 2,500 msec^2, T_{b1} had a mean of 150 msec and variance of 500 msec^2, and T_{b2} had a mean of 250 msec and variance of 1,500 msec^2, then their sums would produce the distributions given above. In the absence of a model, the utility of such a decomposition is minimal, however. It should be easy to see that there exists a large number of RT-type distributions for the variables T_{ij} that will pass the summation test without assuming seriality of components or selective influence. (For example, one need only add exponential deviates of equal rate to each normal distribution T_{ij} to see that the same argument holds for the resulting ex-Gaussians.)

Roberts and Sternberg (1994) also noted that the presence of a third factor, *C*, that influences both process *A* and *B* will not affect the outcome of the mixture test, but might possibly cause the summation test to fail. In two of the data sets they investigated, they observed such an interaction. For example, in one experiment, which we will simulate here, the stimulus materials interacted significantly with the critical factors *A* and *B*. Therefore, they performed a rescaling of the distribution functions of the sums for each level of the interacting factor *C* (stimulus) for the summation test, but not for the mixture test. These rescaled distributions were then averaged over the level of the factor. This rescaling equates the means of the first, second, and third quartiles of the distributions on the left and the right side of the summation test equation. Distribution functions can be highly similar, regardless of their actual shapes, due to their limited domain and monotonicity. Thus, it was not clear to us whether the success of the serial model was due to the rescaling procedure. This turned out not to be the case, as we will demonstrate.

First, to determine the effect that statistical mimicking might have on the summation test and to investigate the potential problem with the rescaling procedure, we simulated the alternate-pathways model mentioned above using binary mixtures of ex-Gaussians. Recall that the alternate-pathways model is a binary mixture of two processes, and the effect of the experimental conditions was to lengthen the duration of one or both processes, but the mixing proportions were unchanged. Second, we examined the behavior of a mixture process (similar to the one discussed during the examination of the IES model) in which five normal distributions contributed to the observed processing time, and the effect of the experimental conditions was to change the proportion of observations observed from each one. For each model, we simulated conditions in a 2 × 2 factorial experiment. The effect of each factor was to decrease the rate parameters of the exponential portion of the distributions in the alternate pathways model and to change the mixing proportions in the mixture model. For the mixture model, it was assumed that the underlying process gave rise to normally distributed processing times, and the mean and variance of these times varied from trial to trial. The effect of increasing the levels of factors *A* and *B* was a change in the distributions of the mean and variance (see Table 2). For this model, factor *A* produced a main effect of 408 msec, and factor *B* produced a main effect of 202 msec. This model predicts a small underadditive interaction effect of −41 msec. This interaction should be sufficient to cause the summation test to fail, but it is small relative to the main effects. It is of interest to determine if the summation test is sensitive to it.

Figure 13 shows the ex-Gaussian densities of the subprocess completion times for the alternate-pathways model (solid lines). The mixture densities produced by the alternate-pathways model are shown with a dotted line. In each panel, the alternate-pathways density was produced by combining .427 of the *A* density and .573 of the *B* density. The effect of increasing the level of factor *A* was to increase the exponential mean from 300 to 400. The effect of increasing the level of factor *B* was to increase the exponential mean from 100 to 200. For all ex-Gaussians, the normal component was held constant with a mean of 100 and a standard deviation of 100. The additivity of the mean RTs are preserved with these parameter values; factor *A* produced a main effect of 43 msec, and factor *B* produced a main effect of 57 msec.

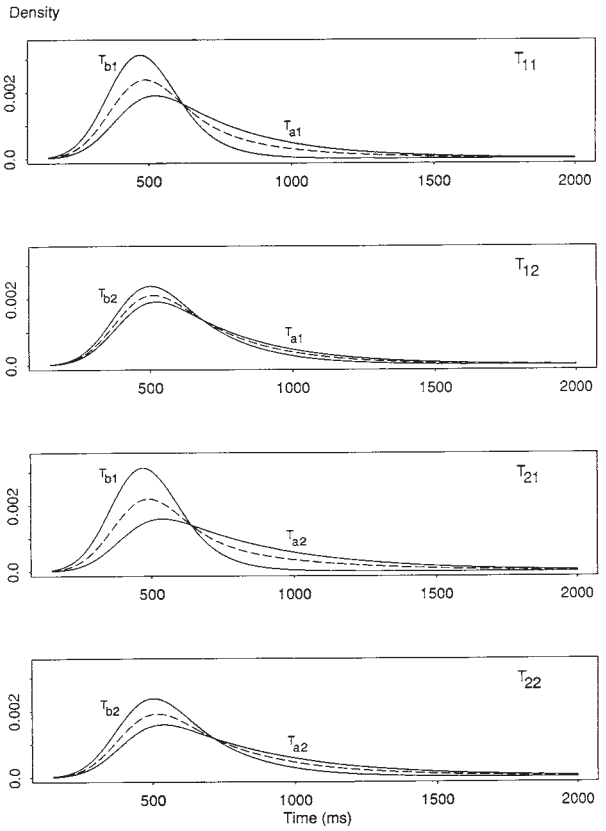

Figure 13. The processing-time densities of the alternate-pathways variables. The densities for each level of factors *A* and *B* are presented as solid lines, and the mixtures are presented as dotted lines. As factors selectively prolong paths *a* and *b*, the exponential component of the ex-Gaussian distribution increases. This is seen as an increase in the variance and a decrease in skew for variables T_{a2} and T_{b2}.

Table 2
Mixing Proportions of the Mixture Model Under the Four Conditions in a 2 × 2 Factorial Experiment

Normal Parameters		Condition			
M	*SD*	*A*=0, *B*=0	*A*=0, *B*=1	*A*=1, *B*=0	*A*=1, *B*=1
400	150	.545	.074	.005	.005
600	180	.195	.426	.129	.005
800	230	.232	.140	.268	.089
1,000	300	.023	.327	.423	.536
1,500	400	.005	.032	.174	.365
Model RTs					
M		551	773	979	1,160
SD		184	240	291	334

Note—Means (*M*s) and standard deviations (*SD*s) are given in milliseconds.

For both models, we simulated the effects of a third interacting factor, *C*. Toward this end, for each of 5 subjects, 40 observations were simulated for each response condition in each of eight levels of *C*. A normal deviate of mean 100 and a standard deviation of 100 was generated and added to all observations at one level of *C*. Between-subjects variation was also simulated by adding, for each of 5 subjects, a normal deviate of mean of 600 and a standard deviation of 100. The number of subjects, levels of *C*, and the number of observations in each level closely replicated the conditions of one of the experiments that Roberts and Sternberg (1994) examined (Sternberg, 1969, Experiment V). Summing the observations in the appropriate combinations for each condition thus yielded 1,600 observations on each side of the equation for the summation test for each stimulus and subject.

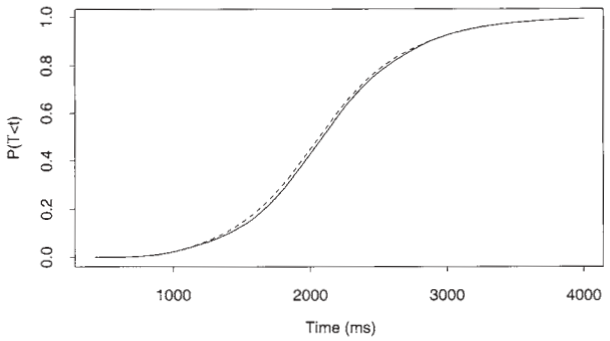

Figure 14. The results of the summation test applied to the alternate-pathway variables shown in Figure 13. Pictured are the average distributions of the sums for the left (solid line) and right (dotted line) sides of the summation test equation.

Analyses of variance (ANOVAs) performed on the simulations of both models showed no significant interaction effects on either the means or the variances, so both simulations could possibly pass the summation test.

Following Roberts and Sternberg's (1994) procedure, the observations were rescaled. The medians and interquartile ranges of the left and right sides of the summation equation were calculated for each stimulus and subject. These were then averaged over the left and right sides of the equation. These averages were then averaged over stimulus and subject. Using the resulting within- and across-stimulus and within- and across-subject averages, all observations were then subjected to the linear transformation presented by Roberts and Sternberg, which equates the means of the medians and interquartile ranges over all conditions on each side of the summation equation. Analyses were then performed on the quantiles and percentiles of the distributions, using the variation between subjects as an estimate of standard error. As an example of the resulting mean distributions, the average summation distribution functions (from Equation 2) for the alternate-pathways model is shown in Figure 14. The agreement between the left and right sides of the summation equation is very good.

For the quantile analyses, we followed Roberts and Sternberg's procedure by examining the quantiles at p values of .05, .10, .25, .50, .75, .90, and .95, as well as the interquartile ranges (IQRs), a measure of skewness ($t_{.95} + t_{.05} - 2t_{.50}$), and a measure of kurtosis [$10(t_{.95} - t_{.05} - IQR)$]. We performed two-tailed t tests on the differences between each quantile measure on the left and right side of Equation 2 for each subject individually and on the grouped data. The variance across stimuli was used as a measure of standard error for the individual analyses, and the variance across subject means was used for the grouped data. For the alternate-pathways model, there were no significant differences between any measure between or within subjects. One subject showed marginal differences for the quantiles at high p values, but none reached significance (all ps > .03). For the mixture model, the only significant difference between the two distributions was in their skewness [$t(4) = -2.92$, $p < .01$]. One subject also showed a significant difference in skew [$t(7) = -3.02$, $p < .02$].

For the analysis of the percentiles, we performed t tests on the differences between the mean distribution functions within and across subjects at 10-msec points along the time axis. Several hundred t tests were performed for each subject. For the alternate-pathways model, only 1 subject showed a single significant deviation between the two distributions. Across subjects, the grouped data showed two significant differences among 298 points. For the mixture model, approximately 350 t tests were performed for each subject. No significant differences were observed. Across subjects, the grouped data showed one significant difference among 360 t tests.

We plotted the average differences between the distribution functions for both models (the alternate-pathways model distributions are shown in Figure 14). These differences, plus and minus one standard error, are shown in Figure 15. The top panel shows the residuals for the alternate-pathways model, and the bottom panel shows the residuals for the mixture model. Both models have passed the summation test quite well. The alternate-pathways model, after rescaling and averaging over subjects and the effects of a confounding variable, passes

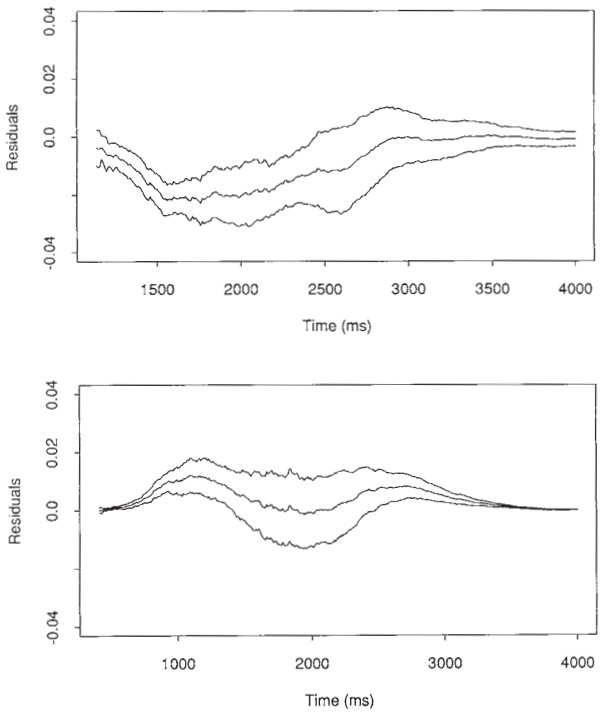

Figure 15. The results of the summation test applied to the distributions of the alternate-pathways model shown in Figure 13 and the mixture model of Table 2. The differences between the average distribution functions for the left and right side of the summation test equation are plotted along with the standard error based on between-subject variation. The central line is the residual, centered between +1 and −1 standard error for the alternate-pathways model (top panel) and the mixture model (bottom panel).

the test for additivity that logically excludes mixture processes. The mixture model, which predicted a small negative interaction, also passed the summation test despite the absence of serial stages. However, the success of these models is not due to the rescaling procedure. When the data were not rescaled, but were subjected again to the same sequence of *t* tests on the quantiles and percentiles, the results of the analyses were unchanged from when the rescaling was performed. What we have observed, then, are simply instances of statistical mimicking. The rescaling procedure (for these data) does not have the effect of increasing the similarity between the summation distributions.

We also subjected the two sets of data to the mixture test, to compare the results with those of the summation test. We expected the data from the alternate-pathways model to pass the mixture test as well as or better than they passed the additivity test. The generation of data from the mixture model violated most of the conditions required for the mixture test to hold, so we were fairly certain that these data would not pass the mixture test. The data from both simulations were rescaled in a fashion similar to that for the additivity test, to equate the average medians and interquartile ranges across the left and right sides of the mixture test equation for all subjects, but the data were collapsed across stimulus without rescaling and were averaged over subjects.

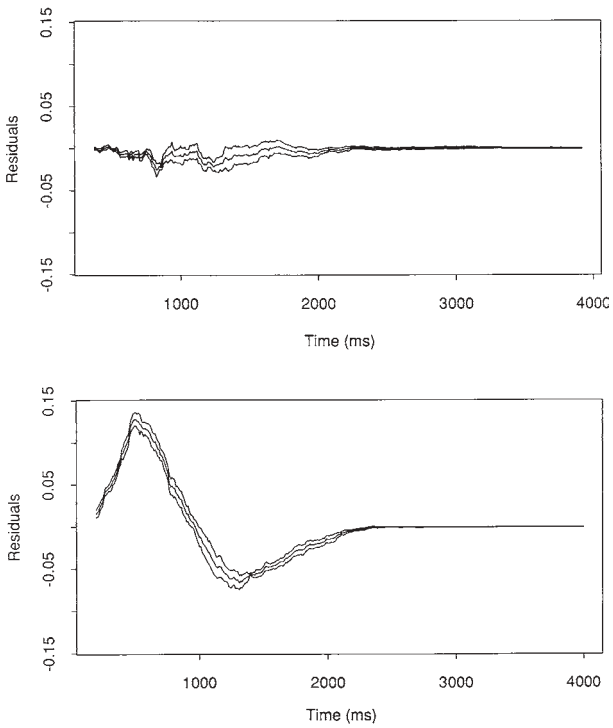

Figure 16. The results of the mixture test applied to the distributions of the alternate-pathways model shown in Figure 13 and the mixture model of Table 2. The differences between the average distribution functions for the left and right side of the mixture test equation are plotted along with the standard error based on between-subject variation. The central line is the residual, centered between +1 and −1 standard error for the alternate-pathways model (top panel) and the mixture model (bottom panel).

Analyses of the quantiles for the alternate-pathways model showed no significant differences between any of the quantile measures we examined. For the mixture model, however, all quantiles were significantly different (largest $p < .005$), except for the estimate of kurtosis. For the analyses of the percentiles, multiple *t* tests performed on the differences between the left and right sides of Equation 3 at each 10-msec interval along the time axis showed 7 significant differences among 355 *t* tests for the alternate-pathways model, and 155 significant differences out of 381 *t* tests for the mixture model. The residuals of the mixture test are presented in Figure 16 for the alternate-pathways and the mixture models. The top panel shows the difference between the left and right sides of the mixture test equation for the alternate-pathways model, and the bottom panel shows the residuals for the mixture model. The standard error bars are calculated from the variation between subjects. The alternate-pathways model has passed the mixture test as well as it passed the summation test, whereas the mixture model has failed the mixture test in both the quantile and the percentile analyses.

For the summation test to be passed, the data must show additivity for both the means and the variances. Sternberg (personal communication, December 1993) has pointed out that one reason for the inability of the summation test to fail for the alternate-pathways model may lie in the relative sizes of the variance to the mean main effects. The alternate-pathways model predicts a variance interaction that is proportional to the product of the mean main effects. Because the model's variance is quite large (88,710 msec^2, calculated as an average across the four conditions) and the mean main effects are small (43 and 57 msec for factors *A* and *B*, respectively), the expected variance interaction was dwarfed; indeed, the variance interaction was not significant in the ANOVA. Thus, the alternate-pathways model shows the mean and variance additivity necessary (but not sufficient) for the summation test to be passed. However, this difficulty is not encountered for the mixture model. The mixture model's variance (70,482 msec^2) is actually smaller than the product of the mean main effects (408 and 202 msec for factors *A* and *B*, respectively). If the summation test was passed by the data from the alternate-pathways model because of the small relative size of the mean main effects, the ability of the mixture model to pass the summation test must be due to some other factor.

Summary

The additive-factors approach to testing hypotheses in psychology has had a powerful influence on the way that cognitive research is conducted. Its extension to the RT distributions in the form of the summation test provides an even more rigorous way to search for serial stages of processing in cognitive tasks. However, just as with the additive-factors approach to the analysis of mean RT, the summation test for serial stages must be applied with

care. It is a valuable tool, but it may be dangerous to use it as a distribution-free test for the presence of serial subprocesses without considering other performance variables and other aspects of the task under study. We have constructed several process architectures unlike the serial subprocess model that can pass the summation test quite well. The important issue for use of this method is *power*. Will it be easy to produce alternative models of the sort presented above, or will each result supporting a serial model be fit only by a different idiosyncratic model? Our limited experience suggests that the truth lies somewhere in the middle of these two extremes.

The problem of statistical mimicking of the IES and summation tests arises due to post hoc interpretation of data and a lack of processing considerations. If we knew why the additional subprocessing time in the IES model was exponentially distributed, or if we knew more about what was happening inside the serial processes proposed for the summation test, perhaps predictions concerning accuracies, confidence, and so on could be made and used to distinguish between the different models. This is not a shortcoming of these tests, but a shortcoming of the way they were applied. So, we again stress the importance of using tests such as these from within the confines of explicitly defined models of the processes of interest, rather than using them to make post hoc assumptions about cognitive structure.

In the next section, we will examine a model of performance that makes distributional predictions about RT that fall naturally from the hypothesized structure of the process. We will demonstrate that the RT predictions can be statistically mimicked by a single-process model; however, because of the clearly defined structure of the model, this is not as great a problem as it is for the IES or the serial stage models.

CASE 3
Instance Theory

With practice, a subject's performance of a task, such as lexical decision, memory search, and so on, becomes "automatic." There are conflicting assumptions proposed by various researchers outlining exactly what is meant by automatic, but, in general, an automatic process is one that requires little or no mental resource for its performance. The task becomes relatively effortless; RTs descend to floor and accuracies are very high. Often, the task is performed without conscious awareness or in a mandatory fashion (Posner, 1982; Schneider & Shiffrin, 1977; Strayer & Kramer, 1990). To explain the acquisition of automatic performance, Logan (1988) has proposed an "instance-based" theory of automatization. He theorized that, over the course of practice with a task, performance gradually shifts from a reliance on the algorithmic solution of the task to the retrieval of the memories for specific instances of a problem and its required solution.

For example, in a memory search task, a particular probe item may require a negative ("not present") response. A subject might search the memory set for this probe item at the first few presentations of it; however, after that, if that item always appears as a probe and never a distractor, the memory for the negative response given to that probe should be sufficient to perform the task. The process that Logan proposes to underlie the shift from algorithm to memory is a race. Each presentation of the probe lays down a separate and independent trace in memory. The retrieval of these traces races with the execution of the algorithm, and the RT is determined by the time taken by the shortest of these processes. As more and more instances are encountered, the likelihood that one of those instances is retrieved before the algorithm is completed becomes higher and higher. The statistics of the race take over; the additional members of the race represented by each new instance cause the winning time to become smaller and smaller. This increase in speed captures the qualitative aspects of automatic performance observed in experiments on skill acquisition, and also the shift from effortful algorithmic problem solving to effortless retrieval of a solution from memory.

Logan (1988) discussed at some length the characteristics of the RT means and standard deviations as functions of the number of presentations of a stimulus. The "power law of practice" is well known (e.g., Newell & Rosenbloom, 1981); it describes the decrease in mean RT and the RT standard deviations as power functions ($b + a\,N^{-c}$) of the number of trials or exposures to a stimulus (N). The parameters a, b, and c are positive; b usually represents the contribution of perceptual and/or response execution processes, and c determines the rate of improvement. Instance theory nicely predicts these relationships as an outcome of the race process. The outcome of the race is modeled as a minimum statistic—that is, the smallest of N random variables is observed. Under general conditions (which simply assure that each instance of a stimulus is similar to the ones that have been laid down earlier; see Fisher & Tippett, 1928), if these observations are positive (as RT data are), then as N grows, the minimum statistic converges in distribution to a Weibull random variable (cf. Colonius, 1993). As the number of instances (N) increases, the number of runners in the race increases, and the observed RTs will tend to be distributed as Weibull random variables, with distribution function

$$F(t|N) = 1 - \exp\left\{-\left[N^{1/c}\left(\frac{t-b}{a}\right)\right]^{c}\right\}. \qquad (4)$$

The mean, standard deviation, and quantiles of the Weibull distribution decrease as a power function of the number of instances, N.

Thus, instance theory predicts that after some number of trials with a particular stimulus–response combination, the distribution of RTs to that combination should be approximately Weibull. Furthermore, the means and standard deviations should follow the power law of practice, with the same exponent, $1/c$. Logan (1988) fit power functions with identical exponents to the RT means and

standard deviations, providing support for instance theory. He also demonstrated, in the course of several experiments, that performance of lexical decision, alphabet arithmetic, and pronunciation decision tasks appears to be item-based—that is, performance depended on whether a particular stimulus–response combination had been encountered before, and transfer to new stimulus–response pairs was poor. Furthermore, transfer to a final frequency judgment task produced stimulus-frequency judgments that were independent of the consistency of the stimulus–response mapping, supporting the notion that each stimulus presentation lays down a new trace in memory.

Because instance theory also predicts that the RT distribution should be asymptotically Weibull and its quantiles decreasing as a power function of the number of stimulus presentations, Logan (1992) examined the entire RT distribution. Using the same exponential constant used to fit power functions to the means and standard deviations, he showed that the Weibull could be well fit to the RT distributions from several tasks. The quantiles of these distributions decreased as N increased. Thus, his RT data are consistent with those predicted by instance theory even at the level of the RT distributions.

However, the Weibull distribution is not the only distribution that predicts a power function decrease in the RT quantiles, means, and standard deviations. In fact, any random variable T that decreases over the course of an experiment by N^{-c} ($T = N^{-c}X$ for some positive random variable X) will also predict a power function decrease with exponent c in mean, standard deviations, and quantiles (see Appendix B). It need not be the case that the RTs are Weibull-distributed; hence, the assumption of a race process may not be necessary. Because the Weibull is one of very many positive, unimodal and positively skewed distributions, its nice fits to RT distributions are not surprising. Many other unimodal and positively skewed distributions might fit as well and exhibit the same power-function decrease in means, standard deviations, and quantiles.

Consider, as an alternative to instance theory, a model in which mean algorithmic solution time for a stimulus decreases as the number of stimulus presentations increases. The algorithm does not change over different levels of practice; it simply gets faster. To construct this model, we can eliminate the laying down of traces (and, hence, the "race" from instance theory) and presume that each stimulus presentation causes the mean processing rate of the algorithm to increase. At the second presentation of a stimulus, the operation of the algorithm is greatly speeded relative to the first presentation. However, at the third and fourth presentations, the amount of relative speed-up is not as great. Eventually, the rate of processing asymptotes at the point where the algorithmic process reaches its physical limitations.

The algorithm might proceed by delivering units of "information" to a response selection process. For a lexical decision task, this information could take the form of the number of connections recovered in semantic memory. Units arrive from the algorithm in exponentially distributed time intervals, and the response selection process waits until some predetermined number of units has accumulated. The time to initiate a response is then distributed as a gamma random variable with a shape parameter equal to the number of units necessary for response initiation and a rate parameter equal to the rate at which the algorithm delivers information to the response process.[4] As the frequency of a particular stimulus increases, the rate at which the algorithm can find and deliver information increases. This might arise because the previously discovered connections are primed or strengthened. Suppose that this rate parameter is also a random variable, varying according to the momentary demands on the algorithmic process and the residual effects of previous stimuli. Let the distribution of the rate parameter also be gamma, with fixed shape parameter and a rate equal to a power function of the number of stimulus presentations. Across trials, as the number of stimulus presentations increases, this model produces observed RTs that are continuous mixtures of gammas, with RT means, standard deviations, and quantiles that decrease as a power function of the number of stimuli with the same exponent (see Appendix C). This model also predicts positively skewed unimodal densities typical of RT data.

We fit this model to some of the data presented by Logan (1992, see his Figure 9) involving an alphabet arithmetic task. In alphabet arithmetic, a subject is presented with a problem such as $A + 5 = F$, and the task is to determine if the letter F is five positions past the letter A in the alphabet. The equation $A + 5 = F$ would require a "true" response, whereas $B + 3 = D$ would require a "false" response. We used two conditions of the alphabet arithmetic data, where the addend equaled 2 or 4; we chose these conditions because instance theory fit the addend = 2 case best and the addend = 4 case worst. For each addend condition, Logan used a single set of parameters for all distributions. There were six levels of practice across the experiment, and the stimuli were presented six times at each level. To fit the mixture model,

Table 3
Parameters and Goodness-of-Fit Values for the Mixture Model and Logan's (1992) Instance Theory

Addend	Correct Response	Goodness of Fit *rmse*	Goodness of Fit r^2	Parameters *a*	*b*	*c*	*k*	*j*
			Mixture					
2	True	90	.994	2,743	562	1.528	10.113	4.773
	False	94	.995	5,390	416	1.759	8.520	7.025
4	True	51	.999	6,612	69	1.644	14.536	9.240
	False	113	.996	8,566	0	1.701	13.124	10.562
			Instance Theory					
2	True	119	.979	4,909	656	1.766	—	—
	False	135	.979	5,534	575	1.984	—	—
4	True	174	.979	7,164	394	1.993	—	—
	False	214	.971	7,426	276	2.123	—	—

Note—Parameters k and j represent shape parameters for the processing-time distribution and the rate distribution, respectively.

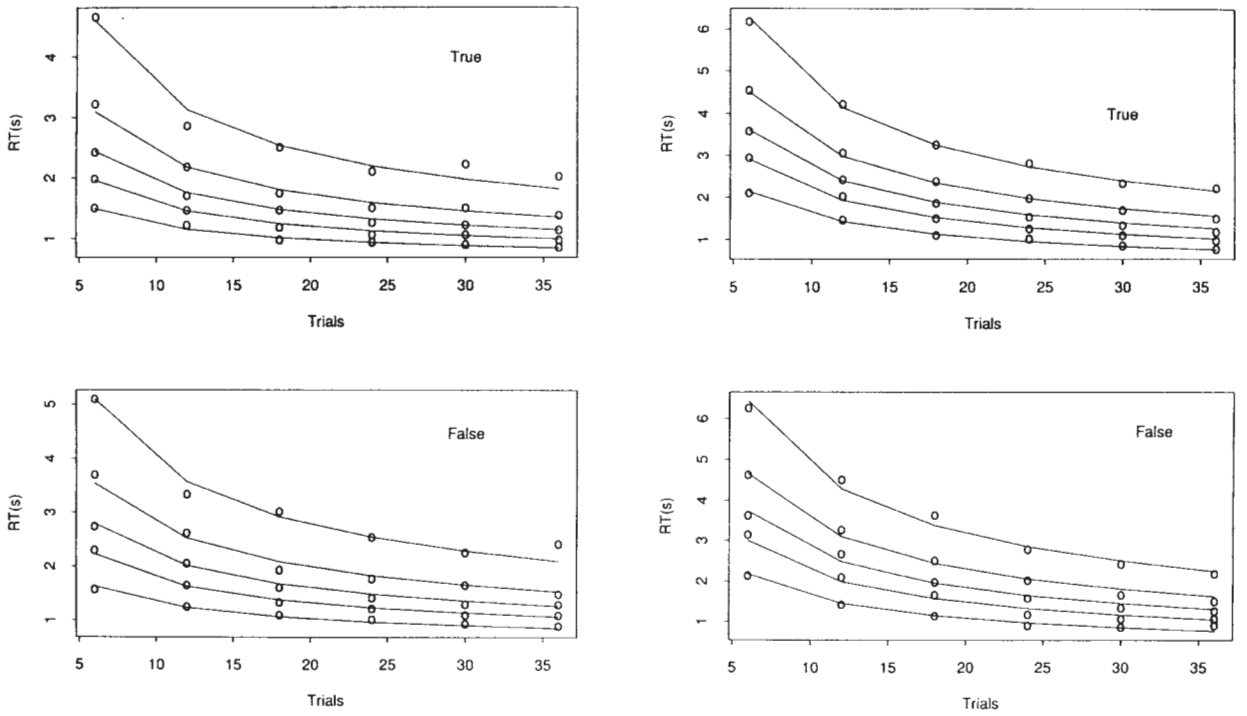

Figure 17. Fits of the mixture model to Logan's (1988) alphabet arithmetic data. The quintiles of each distribution are shown as the open circles at each level of practice. The curves are the predictions of the mixture model. The two left panels show the fits to the case where the addend equaled two, and the two right panels show the fits to the case where the addend equaled four. The top panels show the RTs for the case where the correct response was "true," and the bottom panels show the RTs for the case where the correct response was "false."

five parameters were estimated for both the true and the false conditions and for each addend. These included the three parameters appearing in the power law, including the time intercept and the exponent c as well as the constant a, and the two shape parameters required for the mixture model. The mean rate of information accrual at level of practice N is expressed as $aN^{-1/c}$; for these data, $N = 6$, 12, 18, 24, 30, and 36. All parameter values are presented in Table 3. Goodness-of-fit measures r^2 and root mean squared error (*rmse*) are also presented in Table 3. The fits of the mixture model to the RT quintiles presented by Logan are shown in Figure 17. The top panels show the RTs for alphabet arithmetic equations that were true, and the bottom panels for those that were false. The left panels show the RTs for addend = 2 conditions and the right panels for addend = 4.

The fits of the mixture model are excellent. There are no systematic deviations of the data from the predicted curves. Each curve in each panel was generated by varying N only. The r^2 statistics are higher for the mixture model than for instance theory, and the *rmse* statistics are considerably smaller for the mixture model than for instance theory. This is perhaps not surprising, since there are two additional parameters (k and j) for the mixture model. It could potentially be difficult to distinguish between these two models—one that assumes a single process that gets faster and another that assumes a race between many processes—on the basis of their fits to RT data.

However, there are two aspects of the RT data that instance theory cannot accommodate. The first of these is the behavior of the parameter b, which is assumed to encompass the time required by perceptual and motor processes. Logan demonstrated that the parameter b also decreases as a power function of practice. An examination of Equation 4 shows that there is nothing in the assumptions of instance theory that could explain this decrease. Instance theory, which produces the $N^{1/c}$ term in the equation, predicts only an effect on the scale parameter a, and the addition of more traces should leave the shift parameter b unaffected. As Logan mentions, it may be reasonable to assume that the residual processes embodied in b also speed up as a power function of practice. The second aspect of the data involves the hazard functions. Because instance theory predicts Weibull-distributed RTs, it also predicts that the RT hazard functions should be monotonic increasing (since the parameter c is greater than one). As we mentioned earlier, RT data usually exhibit monotonic increasing or increasing then decreasing hazard functions. As performance improves, the hazard functions should pull up and to the left, becoming nonmonotonic. Instance theory cannot predict this behavior. Because the mixture model is a mixture, however, it can produce the nonmonotonic increasing then decreasing hazard functions typical of RT data (see Figure 18).

As Logan states, the important aspect of his analyses is not that they demonstrate that the Weibull distribution

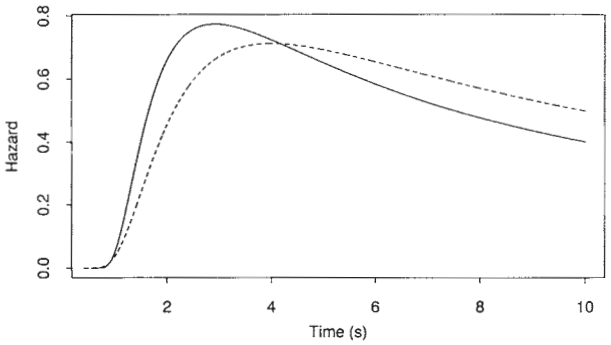

Figure 18. The hazard functions predicted by the mixture model for the addend = 2 condition. The solid line is the "true" RT hazard function, and the dotted line is the "false" RT hazard function.

fits the data but rather that the instance theory naturally predicts the Weibull as a result of its processing assumptions. The same cannot be said of the single-process mixture model. Several assumptions were selected arbitrarily for the single-process model to demonstrate a distribution-level mimicking problem. How can instance theory be modified to produce the decrease in b and a potentially nonmonotonic hazard function? The most straightforward solution is, as Logan himself suggested, to assume that b is a random variable that systematically decreases with practice. Accounting for the variability in b produces a mixture and thus opens the way for instance theory to produce a nonmonotonic hazard function. The hazard function of a mixture of the distribution used to fit the alphabet arithmetic data for addend = 2, $N = 6$ with a varying shift parameter (with exponent equal to .5 and rate equal to .33) is shown in Figure 19. If the rate parameter of the shift distribution is an increasing function of practice, then the mean of the shift parameter will decrease with increasing practice.

We must emphasize again that the single-process model presented here is not intended as a serious competitor to instance theory. The fact that it fits the RT data is inconsequential when considered in light of the lack of foundation for the original distributional assumptions. For instance, why should the rate parameter be gamma distributed? Logan's (1988,1992) approach to modeling automaticity is an instantiation of (in our opinion) how RT data should be used to test models. RT predictions are, first and foremost, model driven; only by assuming the race between traces does the Weibull prediction arise. The race also specifies the behavior of the means and standard deviations over practice, as well as the change in the Weibull rate parameter. Instance theory also makes predictions about other aspects of performance—namely, frequency judgments—and the single-process model does not. The only shortcoming of instance theory that the preceding exercise revealed was that the behavior of the Weibull hazard functions was not consistent with that of typical RT hazard functions. However, this problem can be circumvented by introducing variability into the shift parameter.

If the Weibull distribution had been chosen arbitrarily to represent the finishing-time distribution of a process too vaguely defined to require something else, then the demonstration that the single-process model could also produce the same patterns in the RT data, and fit the data better, would have been a serious blow.

We have concentrated on the entire RT distribution for this example. We will now examine a model of mean RT data and demonstrate that the problem of statistical mimicking is quite serious.

CASE 4
The Guided Visual Search Model and Mean RTs

In a visual search task, an observer is provided with an item, or *target*, and then is presented with a visual array consisting of several items. The observer's task is to respond positively if the target is present in the visual array and to respond negatively otherwise. It has become common to use stimuli that can be described by values on particular stimulus dimensions, such as color, shape, and size. Suppose, for instance, that an observer is looking for a red X in an array that might contain red and green Xs and Os. If all of the nontargets in the array are green or Os, the observer can respond very quickly as to the presence or absence of the target. However, if the distractors are red Os and green Xs, the observer's response is much slower and much more error-prone. This type of search task is called *conjunction search*, since the observer must discriminate between combinations of values on the two possible stimulus dimensions. Treisman and Gelade (1980) proposed a two-stage model of visual search, in which a fast parallel preattentive stage could sort stimuli in the array on the basis of the stimulus dimensions. If the target is red, for example, and all nontarget items in the display are green, the preattentive stage can quickly classify all the display items by color, and the response can be executed immediately. If the target is a conjunction of the stimulus dimensions, the

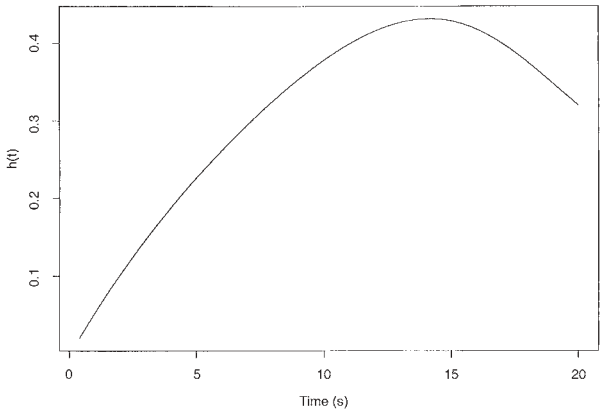

Figure 19. The hazard function predicted by instance theory with a Weibull-distributed shift parameter. Parameter values are given in the text.

preattentive process fails, and a slower serial search is executed, in which each item in the display is examined.

Wolfe et al. (1989) presented results that seemed to be inconsistent with Treisman and Gelade's (1980) two-stage model of visual search in that information from the preattentive stage of processing appeared to be eliminating distractors for comparison during the subsequent search stage. Using easily discriminable feature values for each dimension, they found that RTs to conjunction targets were much faster than those reported by Treisman and Gelade and that the slopes of the RTs as a function of display size were much shallower. Thus, the "serial" process that they observed did not appear to be as "serial" as that observed by Treisman and Gelade. Moreover, they showed that when three stimulus dimensions were involved, triple conjunction targets were easier to discriminate than were simple conjunction targets— a finding to which Treisman's feature integration model cannot speak.

Wolfe et al. (1989) theorized that the parallel stage of processing produces a spatiotopic attention map of the visual display. Activation is directed to each location in the attention map that corresponds to the feature value of the target. Thus, more activation is given to the locations where features conjoin. Using these activation levels as guides, a serial search ensues, where attention is drawn to the highest activation levels in the map first, and the search terminates when the target is found. The guided search model predicts that multiple conjunctions will be easier to detect than will simple conjunctions among distractors that share one feature with the target, because the additional feature dimension will increase the activation level of a particular display location over that of simple conjunction stimuli. When the distractors are themselves conjunctions, sharing two features with the target, then triple conjunctions should be slowed again as in regular conjunction search.

Wolfe (1994) has recently elaborated the theory behind the construction of the attention map, which now allows the guided search model to account for perceptual grouping phenomena, among other things. This expanded theory allows changes in the peaks of activation in the attention map based on the organization of the visual display and the observer's expectancies. Once the map is constructed, however, search of the display ensues as before, using the activation levels as guides.

The attention map actually contains a great deal of information, and this information may be sufficient for the selection of a response (Pavel, 1990). If one location has a much higher activation level than the others, it is very likely that a target is located in that position. Thus, target displays will have higher average activation levels than nontarget displays. A search process is not actually necessary, since the response could be based on this perceived level of overall activation. To demonstrate this, the subprocesses responsible for search can be eliminated and the "present" or "absent" judgment can be made on the basis of the activation levels alone (e.g., Pavel, 1990). After a little practice with the task, an observer gets a feel for the average activation level across the surface of the attention map when a target is present versus when a target is not present. The observer sets a criterion along the continuum of mean activation that is perceived and tends to respond negatively when the perceived mean activation is less than the criterion and positively otherwise.

More rigorously, for a given level of stimulus clarity, target discriminability, and so on, each stimulus contributes a random amount of activation to its position in the attention map for each feature that it shares with the target. For each dimension and each stimulus the amounts contributed are independent and identically distributed with some mean, u. If there are two stimulus dimensions for a simple conjunction search, in which the distractors each share one feature with the target, and N locations in the attention map, then the average activation level for each position will be u for nontarget trials and $[(N+1)/N]u$ for target trials. The measure of discriminability d' (under the assumption of equal variance for the target and nontarget activation distributions) is proportional to $[(N+1)/N]u - u = u/N$ for these trials. For triple conjunction search among nontargets that share a single feature with the target, the nontarget distribution of perceived activation is unchanged, but the target distribution then has a mean of $[(N+2)/N]u$, and d' doubles, being now proportional to $2u/N$.

For the condition where the distractors of a triple conjunction search share two features with the target, the noise distribution is then shifted upward to have a mean of $2u$, and d' again drops to the level of simple conjunction search. For each of these conditions, simple conjunction, triple conjunction, and triple conjunction with conjunction distractors, the observer sets an appropriate criterion level. In the triple conjunction case where nontargets share only a single feature with the target, perceived levels of activation fall, on average, farther from the criterion because of the increased distance between the target and nontarget distributions. Thus, triple conjunctions should be faster and less error-prone than simple conjunctions. When nontargets share two features with the target, d' is equal to that of the single conjunction case; therefore, little performance difference should be observed.

For all three cases, d' is a decreasing function of N. Therefore, as the array size increases, activation levels will tend to be selected closer and closer to the criterion and RTs will increase. If stimuli are made less discriminable, then the mean amount of activation contributed by each dimension u decreases, decreasing d' and again slowing RTs. It is not necessary to suppose that activation is allocated in the same way for all feature dimensions, but, for the purposes of this discussion, that consideration would provide an unwelcome complication. The constant mean activation model will prove sufficient for now.

It is not necessary to specify the distribution of finishing times produced by the model, since we can always find a system of distributions that can produce the means. To derive predictions for mean RT, however, we

have assumed that RT is a function of the distance from criterion (c) that an activation level (a) is sampled. In particular, we have assumed that

$$E[RT_j|a,c] = \gamma e^{-\lambda_j|a-c|}+b,$$

where the subscript j indicates the asymmetry between positive and negative responses. This function gives a reasonable approximation of the decrease in mean RT as stimulus intensity falls farther and farther from criterion (e.g., Ashby, Boynton, & Lee, 1994; Baddeley & Ecob, 1973; Gescheider, Wright, Weber, Kirchner, & Milligan, 1969; Murdock, 1985). The derivation of the mean RTs based on this assumption are presented in Appendix D.

We made several additional assumptions to reduce the number of free parameters of this model. First, we assumed that the mean activation added by a single feature was one and that its standard deviation was equal to .0075. This standard deviation was selected so that the error rate in the most difficult condition ($N = 32$, triple conjunction sharing 2 features) was around 5%. Second, the amounts of activation contributed by each stimulus dimension to a single location were independent with equal mean and variance. Third, we assumed that the observer was unbiased. These assumptions allowed us to calculate the variance of the activation distributions and the criteria as functions of the set size (see Appendix D), and reduce the parameters to the base times in each condition (b in the equation given above), the decay rates (λ_+ and λ_-), and γ—a total of six parameters.

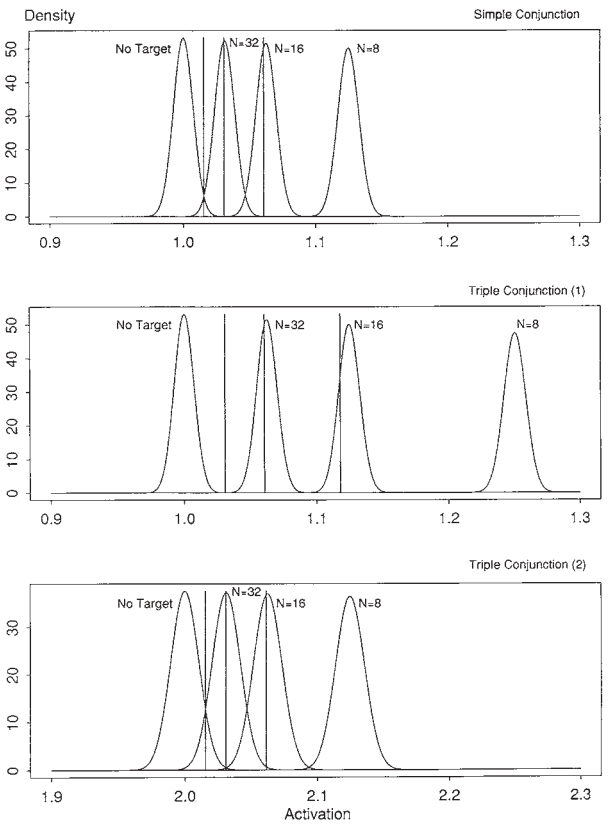

Figure 20. The density of mean activation levels as predicted by the single-process visual search model. The vertical lines in each panel are the unbiased criteria for the $N = 8$, 16, and 32 set sizes moving from right to left. The top panel shows the densities for simple conjunction search. The center panel shows the densities for triple conjunction search where the target and distractors share a single feature. The bottom panel shows the densities for triple conjunction search where the target and distractors share two features. Accuracy should be nearly perfect in the Triple (1) condition, because there is very little overlap between the densities. Accuracy in the Triple (2) condition should be less than in the simple condition, because overlap is greater (notice the difference in scales).

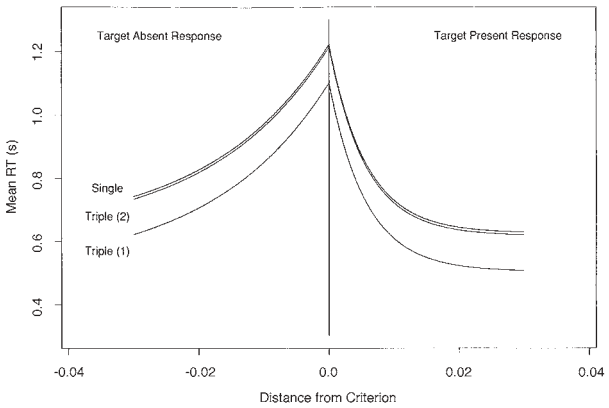

Figure 21. Mean RTs as a function of the distance that a perceived activation level falls from the criterion for the simple, Triple (1) and Triple (2) conditions.

The distributions of mean activation under these assumptions are presented in Figure 20 for all conditions of the conjunction search experiment. The vertical lines in each panel are the unbiased criteria for Set Sizes 8, 16, and 32 moving from right to left along the activation scale. The mean RTs as a function of perceived activation and the six estimated parameters are presented in Figure 21. The only difference between the six functions in Figure 21 is in the base times, the smallest of which is for the triple conjunction condition where distractors share only one dimension with the target [Triple (1)]. The fits of the single-process means to Wolfe et al.'s (1989) data is presented in Figure 22. The model has captured the patterns of mean RT over the different conditions quite well, accounting for 93% of the variance and producing a *rmse* of 31 msec. The error bars, as estimated from a simulation of the single-process model, are superimposed on the predicted means. The standard deviation of the mean decreases as set size decreases because the RTs are at or near floor for $N = 8$. There are two shortcomings in the fits: first, the $N = 8$ base times are too high, resulting in an overestimation of the target-present RT; second, the negative slope in the triple conjunction condition is too large (5 msec) relative to the positive slope (0.3 msec) and the observed slope (1.2 msec). The predictions of the original guided search model are not explicit; therefore, despite these small deviations of the single-process model from the data, comparisons between the two models are not possible.

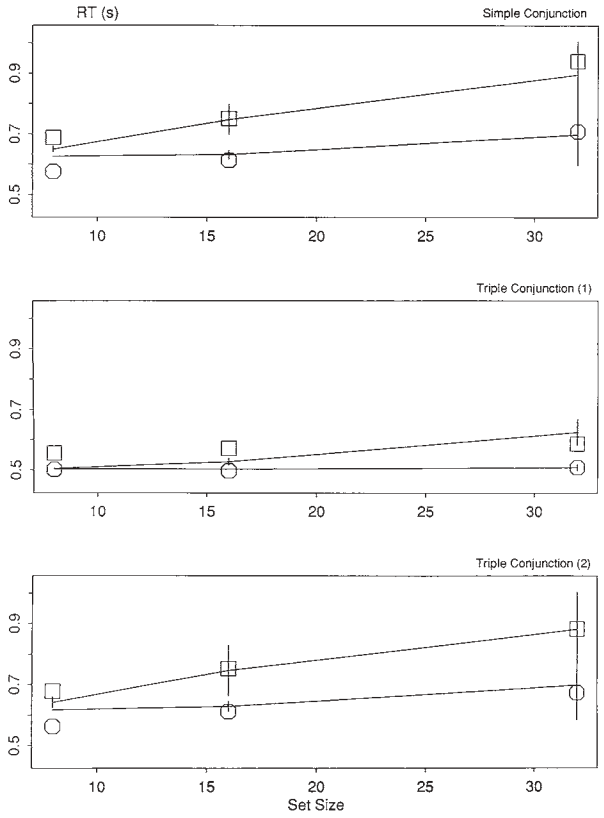

Figure 22. Fits of the single-process model (curves) to the mean RTs (points) collected by Wolfe et al. (1989). The length of the error bars are given by the standard deviation of the mean RT as estimated from a simulation of the single-process model.

The single-process model makes a number of other predictions. The predicted accuracies, for instance, can be explicitly calculated. Furthermore, because of the steady increase in activation variance with increasing set size, although the RT predictions for the simple and triple conjunction search with two shared dimensions may be very similar, the accuracy for the triple condition must be less than the accuracy for the simple condition. This can be seen as a greater overlap between the activation densities in Figure 16 (bottom panel). The model also predicts that the accuracy should be greatest in the triple conjunction condition where only one dimension is shared by the target and the distractors. This demonstration illustrates the importance of considering other variables along with mean RT. Because there exist any number of models that can replicate a given pattern of mean RT data, the observation that, say, mean RT increases with some independent variable does not provide strong support for a model of the process that produced the mean RTs. Also, the RT–distance assumption predicts that error reaction time distributions are negatively skewed, which is inconsistent with RT data, whereas the serial search model is mute on the subject. Observations of this type provide additional tests of the models. They also provide sources of converging evidence, but they are only convincing when the larger patterns of behavior are observed.

In the case of the original guided search model, the mechanisms by which the attention map is formed and the subsequent visual search is carried out were not specified in enough detail to rule out the operation of alternative models in which no search is performed at all. The more recent version of guided search (Wolfe, 1994) is a positive step toward deeper understanding of the processes involved in the performance of visual search tasks. Because Guided Search 2 has made accuracy performance contingent upon the search process, it can be contrasted to models, such as the one presented here, that do not involve search per se. Of course, some structure must give rise to the mean RT–distance function, and "search" may or may not be as good as any other theory. A resolution of the issue will require collecting data with experiments designed to test specific points of the two models.

GENERAL DISCUSSION AND CONCLUSIONS

In this paper, we have examined specific distributional and mean RT predictions made by various established models of performance and shown that, for each of these models, there exist one or more alternative models that also make these predictions. We concentrated on RT distributions. We have shown that even very powerful tests of cognitive architecture based on RT can be passed, or statistically mimicked, by allowing the parameters of the models to vary. This suggests that these tests cannot be applied in a post hoc way and should be applied to a richer constellation of data, of which RT is only a subset.

We began with Ashby and Townsend's (1980) test for the presence of an inserted serial exponential stage of processing, the IES model. This test was passed by several single-process models that did not include a serial exponential stage by changing the parameter values in reasonable ways across experimental conditions. We also presented a model that predicted normally distributed RTs, with randomness in the mean and variance from trial to trial, and that passed the IES test. The RTs produced by this model were mixtures of normal distributions, and the mixture properties rendered the hazard functions ineffectual for distinguishing between the different models. By virtue of the restricted nature of the IES test, it was the most rigorous proving ground for the alternative models. The IES model should have been the most difficult to mimic, but it turned out to be relatively easy to do so. Thus, our ability to mimic other (perhaps more complicated but less rigorously specified) models with alternative representations was almost assured.

We turned from the IES model to the more general serial stage model. The existence of serial stages in mental processing has been addressed using additive-factors logic (Sternberg, 1969). The idea that patterns of interaction between different experimental conditions provides information about the processes underlying the performance of a task has had a tremendous impact on

the way that hypotheses are constructed and tested in psychology and has led to important work on more complicated mental structures (e.g., Fisher & Goldstein, 1983; Schweickert, 1978). Additive-factors logic was expanded from mean RT to the scope of the RT distributions by Ashby and Townsend (1980) and was subsequently applied to a large collection of data sets by Roberts and Sternberg (1994) in the form of the summation test. We presented data simulated from two models that passed the summation test for the presence of serial stages, although neither model included serial stages.

The models that we examined as alternatives to the IES and serial stage models were fit to distributions that we knew would pass the IES and summation tests. In most cases, the parameters of these models were selected in such a way as to minimize the sums of squared error between the distribution functions of the models and gamma distributions. Because the gamma distributions were composed of additive exponential stages, the ability of these models to mimic the various tests was a function of their goodness of fit to the gamma distributions. Although it was not difficult to achieve these fits, we are unable to specify more general conditions under which alternative models will pass or fail the tests. However, Ratcliff (1988) fit the diffusion model to several sets of data and, using those parameter values, the resulting RT distributions passed the IES test for the presence of a serial exponential stage. This suggests that the mimicking problems encountered with post hoc application of these tests could be widespread in the context of nontrivial processing models.

Our findings do not indicate any weaknesses in the IES and summation tests. The problem of mimicking by other nonserial models arises only as a result of the way that the tests were applied. If we were dealing with an explicit model of the processes involved in the separate stages, we could derive predictions for the finishing times of those stages and perhaps examine also the predictions of other variables, such as accuracy or confidence, and their behaviors as a function of other independent variables. Fits of the predicted distributions under the appropriate parameter restrictions, together with the IES and summation tests, would provide strong evidence that, for example, the encoding process and the comparison process in memory search were arranged serially. Other facets of the data could also be examined, and strong empirical manipulations could be made to reveal the limitations of the model under scrutiny. Working backward from the RT data, showing that the IES or summation test is passed, without specifying from a model how the subprocessing times are distributed or how these times interact with other variables, is not convincing by itself since many sets of distributions can pass these tests (at least statistically) without serial components.

As an example of the constraints that an explicit model imposes, we discussed Logan's (1988) instance theory of automaticity. Instance theory accounts for the acquisition of automatic performance of a task via a qualitative shift in processing strategy from the algorithmic solution of a task to a reliance on the memory of its solution. We showed that an alternative model that assumed only that the algorithmic performance got faster with practice could account for the RT distributions as well as or better than instance theory. It was not necessary to suppose that multiple processes were invoked by the memory system or that new memory traces were constructed with every exposure to a problem to account for the RT distributions. As we emphasized, however, the single-process model that we constructed is actually quite inadequate since it accounts only for RTs and only by way of some unfounded assumptions about the algorithm. Instance theory predicts other aspects of skill acquisition, and its prediction of Weibull-distributed RTs is derived from the structure of the model.

Perhaps there are some ad hoc assumptions that could be made with respect to the single-process model to allow it to accommodate other aspects of the data, such as accuracy or frequency judgments; however, without a better model of how the algorithm improves efficiency, instance theory will remain the more attractive option. The RT predictions of instance theory are a direct consequence of the presumed structure of the cognitive system involved, not a collection of convenient mathematical assumptions, as is the single-process model. The point is that, if only RT is considered, these two models are indistinguishable even if the entire RT distribution is examined.

We examined another model that makes predictions about mean RTs in visual search. The potential for statistical mimicking is even greater in the case of mean RT data because there exist any number of models that can produce a set of mean RT data, and it is not even necessary to specify the distribution function of the process producing the data. We emphasized this point by using a signal detection model of visual search, fitting the data that Wolfe et al. (1988) used as support for a multiple-process model of visual search.

Single-Process Models

In many instances, the alternative models that we proposed contained processing architectures that did not change across experimental conditions. Instead, the effect of a change in the experimental situation was to change some aspect of the distribution of a parameter. This is in contrast to the competing models that often rely upon a change in the structure of the central process to account for the RTs: the IES model constructs a new exponentially distributed subprocess, Logan's (1988) instance theory assumes an elaborate memory structure that changes with every new stimulus presentation, and Wolfe et al.'s (1988) visual search model relies upon a multistage search process following the construction of the attention map. What this illustrates is that single-process models have often been neglected when various multiple-process models have been tested against each other.

We can pose the distinction between a multiple- and a single-subprocess model in mathematical terms. Does a change in the experimental situation require either the

specification of a new random variable to represent the duration of a new subprocess or stage (the multiple-subprocess case) or a change in the parameters that determine the finishing time (the single-subprocess case)? It should be clear that it will be very difficult to determine (on the basis of RT data alone) whether an experimental manipulation creates a new subprocess or merely changes the preexisting parameter values of the old subprocess. There will not always be a testable dichotomy between single- and multiple-subprocess models; in some cases, the dichotomy may be purely descriptive. Consider the fact that sums of normal deviates are themselves normally distributed. Thus, a model that assumes several serial subprocesses, each one producing normally distributed processing times, will predict normally distributed finishing times. A change in the mean and variance of a normal distribution from one experimental condition to another is consistent with both a shift in the mean and variance and the addition of one or more stages with normally distributed processing times.

More generally, Townsend and Schweickert (1989, Theorem 1) have pointed out that any experimental manipulation that orders the distributions of pre- and postmanipulation RT is also consistent with the addition of a positive random variable to the premanipulation RT. In other words, if, for every point in time t, the probability of observing an RT smaller than t is always greatest in the premanipulation condition (so the RTs in the premanipulation condition are more likely to be smaller than those in the postmanipulation condition), then we can express the postmanipulation time as the premanipulation time plus some positive random variable (that is not independent of the others). Therefore, any manipulation that (for example) decreases the scale parameter of a single-process model (which is equivalent to making the times longer) can also be mimicked perfectly by a multiple-process model in which an additional (dependent) stage of processing is added (see Appendix E).

What, then, is the point of making the distinction between multiple- and single-process models? Our goal as researchers relying heavily on the observation of RTs is to learn about the fundamental architecture of cognition. It is therefore very important to determine whether changes in RTs across experimental conditions can allow us to infer changes in the structure of the processes in a task. We have shown in this paper that RTs alone cannot be used to make this distinction in a post hoc way. This is because single- and multiple-process models can statistically mimic each other. If we allow for parameter variability, the mimicking problem becomes worse.

One might also ask, in the same spirit, what is the point of making the distinction between serial and parallel processing models, since there exist equivalent serial models to a large number of parallel models? First, it must be realized that there also exist classes of serial and parallel models that do not mimic each other, just as there exist single-process models that do not mimic multiple-process models. We can distinguish between serial and parallel models, as we can distinguish between multiple- and single-process models, by generating a priori predictions about RT and other behavioral variables within different modeling schemes. We can also use more elaborate experimental paradigms that vary several factors and thus observe a far wider range of effects in the data (e.g., Schweikert & Townsend, 1989). This is our resolution to the issue of this paper: testing between different cognitive architectures requires the generation of a priori predictions within a modeling context. The tests that we have discussed are very powerful within the framework of testing the explicitly defined serial models, but single-process models must also be considered.

The single-process model has also been the subject of intense scrutiny recently, and attention has been drawn to the ability of certain single-subprocess models to mimic each other. The models in question are generally classified as accumulator models, in which the response–selection process is guided by the actions of "counters," or neural mechanisms that keep track of the information accruing toward each response. The issue of interest with this type of model is the relationships between the counters. The random-walk model, for instance, can be represented as an accumulator mechanism with counters that are perfectly and negatively correlated; a positive amount of information toward one response decrements an equal amount of information away from all others. Another representation assumes that the counters are independent from each other and that a response is selected as soon as the level on that response's counter exceeds a certain predefined level. Such representations are usually called race models, since the counter that exceeds criterion first "wins" and determines the response.

There has been considerable discussion on the relative merits of the correlated and independent representations, particularly with respect to the kinds of data that each can accommodate (Ratcliff, 1978; Smith & Vickers, 1988; St. James & Eriksen, 1991; Van Zandt, Colonius, & Proctor, 1995; Vickers, 1979; Vickers, Caudrey, & Willson, 1971). For instance, Marley and Colonius (1992) addressed the issue of counter independence directly and showed that any single set of choice RT and accuracy data could be represented by a race between independent counters. In particular, they proved that an equivalent race model exists for any model with counters correlated to an arbitrary degree. Thus, Marley and Colonius assured the existence of statistical, if not exact, race model mimics to random-walk models for choice RT tasks. Dzhafarov (1993) has recently expanded on these results. He demonstrated that any set of RT data can be modeled using what he has termed a "Grice representation." Grice (1968, 1972) presented an accumulator model in which a deterministic counter accumulated evidence toward a random criterion. As with the race model where the task is to select from among several alternative responses, in Grice's model one or more deterministic counter levels race toward variable criteria. Dzhafarov first proved the mathematical equivalence between Grice's model and the stochastic accumulation and deterministic criterion of the race model. Under condi-

tions so general that most RT modelers need not worry about them, he then proved that Grice's model is so flexible as to be "a descriptive language, not an empirically testable model."

Perhaps the best way to understand Dzhafarov's (1993) result is to consider the nature of the distribution functions involved in the Grice model. Define $F(t)$ to be the distribution function of the RTs or the probability that an RT is less than some time t: $P(RT < t)$. Also define $G(c)$ to be the distribution function of the criterion or the probability that the criterion value is less than c: $P(C < c)$. The activation level of the counter at any time t is given by the monotonic nondecreasing function $A(t)$. It should be clear, then, that the event $\{RT < t\}$ is equivalent to the event $\{C < A(t)\}$. Therefore, the probability that an RT is less than t ($F(t)$) is equal to the probability that the criterion is less than the value of the activation function at time t, or $A(t)$ ($G[A(t)]$). The question at hand is can we find an $A(t)$ and $G(c)$ such that the Grice model can produce a particular observed distribution $F(t) = G[A(t)]$? The answer is yes, as Dzhafarov has shown. The relationship between $F(t)$ and $G(c)$ is roughly equivalent to the relationship between a random variable's distribution function and its density function. The behavior of the variable can be characterized using either the distribution or the density (in the case where both exist), and choosing a particular distribution determines the density, and vice versa. For some RT distribution $F(t)$, the choice of criterion distribution $G(c)$ determines the activation function $A(t)$.

Any single set of choice RT data can be modeled using the Grice representation. By inference, any single set of choice RT data can be modeled using a race model or independent counter representation, since Dzhafarov (1993) demonstrated the mathematical equivalence of these types of models. Furthermore, the Marley and Colonius (1992) results guarantee the existence of statistical mimics between independent counter and random-walk models for any realizable set of choice RT and accuracy data collected in a single experimental condition.

The Grice-representability result is different from the serial/parallel equivalence result presented by Townsend (1972). The serial and parallel models specify the structure of the processes underlying RTs, whereas Grice representability is independent from any structural assumptions. Rewriting the RT distribution predicted by a particular process model as a Grice accumulation process does not add to or elaborate upon the predictions of the model. For instance, the standard serial model predicts RTs that are asymptotically normal as the number of serial components increases. Conversely, the independent parallel model predicts RTs that are asymptotically Weibull, if the fastest of the underlying components determines RT. No such dichotomy of predictions need exist for a model and its equivalent Grice representation, because that representation need not be the same over different experimental conditions.

As we stated earlier in this article, questions concerning the presence or absence of parameter variability or the number of processes involved in performing a task cannot be answered within the context of a single experimental condition. This is also the case for accumulator models. Across different experimental conditions, the race model mimics to the random-walk models, and the Grice representations of the RTs under each condition are not constrained to have parameters that are changing in meaningful ways. The parameters required for each representation may not be tied to any reasonable psychological variables, and, hence, the way that they change in order to fit the data may not be readily interpretable. To distinguish between these different classes of models, the parameters must be tied to other performance variables, such as accuracy, confidence, and so on, as well as specific aspects of a well-defined model of a cognitive system. Given such a system, the relationships between the parameters of the equations and the behavior of the process are obvious, and the model can then be tested. Given such a system, dependent variables other than RT manifest themselves, providing alternative ways to test the model. Without such a model, even if equations exist that completely capture the behavior of the RT distributions, those equations may not provide the desired information about the underlying process.

Parameter Variability

We have paid some attention to the notion that experimental data arise from mixtures of processes conditioned on variable parameters. This idea is important because, for a number of reasons, parameters vary over the course of an experiment. At the very least, because the presentation of a stimulus evokes a random perceived effect, the information upon which a cognitive process operates is slightly different for each trial, even if the stimulus presented is the same. As another example, in a memory paradigm, all items encoded into memory are not likely to have the same "strength," so that some variability across items is necessary. Ratcliff (1978) included this assumption into the diffusion model of memory retrieval by specifying the average drift rate as a random variable with a mean held constant within a trial, but variable (with standard deviation η) between trials. In terms of experimental data, in choice RT paradigms there are strong sequential effects that arise as a function of the prior response and stimulus. This means that response time and accuracy averaged across a block of trials incorporate some degree of parameter variability.

The presence of parameter variability implies that RT data are to some extent composed of mixtures of distributions, which render the RT hazard functions more ambiguous than was previously thought. The shapes of the hazard functions of mixture distributions are not constrained by the shapes of the hazard functions of the individual distributions of the mixture (unless the individual hazard functions are all nonincreasing), so typical RT hazard functions (both increasing and increasing then decreasing) can easily arise from different mixtures of the same processes. For example, although the hazard functions of the gamma and inverse-normal distributions

look very different and so could be used as a basis for discriminating between them in the absence of parameter variability, mixtures of gammas and mixtures of inverse-normals can have very similar hazard functions. In sum, the situations where mimicking of RT data by alternative models is a significant concern may be more numerous than it appears.

Conclusions

Suppose that we discover, for a particular task, that RTs follow some distribution F. The distribution F is not merely a statistical estimate of the distribution, but "truth." This knowledge may not allow us to infer as much as we would like about the underlying process in the absence of a model, because there are an infinite number of models that incorporate parameter variability that would produce exactly the observed distribution (see Appendix F). Dzhafarov, a reviewer of this paper, noted quite correctly that this leads one into a situation of infinite regress. If RTs follow some distribution $F(t|A)$, and the parameter A follows some other distribution $G(a|b)$, then the model is complete. The RTs depend on the invariant parameter b, and the problem of variability disappears. However, the parameter b need not be invariant; perhaps it too is a random variable. And so on ad infinitum. Thus, we can always insist that the failure of a model is due not to any shortcoming of the model but rather to some unexplained source of parameter variability. We could potentially justify even the most unlikely model by appealing to enough varying parameters.

Given that subjects adjust criteria, vary their processing rates, or tire of the task over the trials of an experiment, how can we deal with this issue practically from a modeling perspective? The experimental data and the theory must work together to provide an evaluation of the assumptions needed to fit the data. For instance, with the diffusion model, the size of the effects due to parameter variability can be determined systematically. First, blocks of trials can be examined within experimental sessions. As Burbeck and Luce (1982) have outlined, this is one way that the extent of parameter drift (or the range of parameter variability) can be determined between blocks. In several sets of RT data that we have examined, mean RT differences between sessions could be as large as three standard deviations of the mean within sessions. This provides a way to estimate the necessary variance in the mean drift rate. Next, an examination of sequential effects in the data can also indicate the possible extent of criterion variation. In sum, we can attack systematically the problem of parameter variability in the same way that we attack the problem of statistical mimicking: diagnostics that are carried out on thoughtfully constructed model structures and reasoned acknowledgments of sources of parameter variability. Parameter variability is an aspect of modeling that needs to be addressed, and, in the absence of constraints, there is the problem of infinite regress. However, with constraints provided by the data and the theory, the issue is tractable.

In addition, one can find solutions to the problem in the statistical literature. Once a model of the process is specified that predicts the distribution of RTs, then there are several ways available to determine if the observed RT distribution is or is not consistent with that model. For instance, assuming that the model predicts the same family of distributions as that observed, parameter variability can be investigated by looking at the differences between the model and the observed distributions. If the data arises from a mixture of the distributions produced by the model, then the tails of the empirical density function will, in general, be higher than those of the theoretical density function. This observation forms the basis of a test for the presence of mixtures based on the residual differences between empirical and theoretical density functions (e.g., Lindsay & Roeder, 1992).

If a model predicts a distribution arising from a different family from the observed distribution, we must be concerned with the number of ways that a mixture could be constructed from the model distribution to produce the data. This is a question of identifiability, addressed by Tallis (1969). He outlined the conditions under which, for a given model distribution, a parameter density exists for the mixture problem. In the case where no solution exists, the model must be revised. In the case where a single identifiable solution exists, then that particular type of parameter variability can be modeled and tested. However, the conditions under which an infinite number of solutions exist are quite general, and we are again left with a problem of indeterminacy. By working from within the confines of a model, however, various models can be tested and refined. And, just as we are limited in our choice of models by the soundness of their structures, so we are limited in the types of parameter variability with which we might be faced.

The RT measure is important for hypothesis testing and an extremely valuable source of converging evidence for modeling, especially in light of the vast array of diagnostic tests developed for RTs by quantitative psychologists over the past several decades. Used in conjunction with other dependent variables, these tests are extremely powerful tools. However, they must be applied with an understanding of how to avoid potential mimicking by other kinds of models. Using these tools by working backward from the data, attempting to construct viable models on the basis of the tests that the data pass or fail, significantly reduces the power of the tests. Furthermore, no model can be adequately constructed or tested using only a small set of RT data. A larger range of independent variables should be manipulated and, better yet, other behavioral variables should be measured and predefined models applied to all jointly.

It is a risk, we suppose, that those researchers who already carry a bias against quantitative methods will read into our analyses the message that quantitative investi-

gations are fruitless. This is certainly wrong. Cognitive psychology has benefited greatly from the concentrated efforts of quantitative research, research focused on the identification of cognitive structures through rigorous analysis of data and application of mathematically precise logic. If cognitive psychology as a discipline enjoys a more "hard" scientific standing than other areas of psychology, it is due in large part to these efforts. Certainly less quantitative areas of experimental psychology suffer, to an even greater extent, from the problems that we have outlined here, even if the favored dependent measure in these areas is not RT. Our criticism here is directed toward the study of one dependent measure (RT) to the exclusion of others and the collection of RT data without benefit of a model to discipline our thinking. Working from the perspective of a well-defined model of the process of interest, collecting data from experiments designed to test specific points of the model will allow for real progress toward understanding cognitive architecture.

REFERENCES

Ashby, F. G. (1982). Deriving exact predictions from the cascade model. *Psychological Review*, **89**, 599-607.

Ashby, F. G., Boynton, G., & Lee, W. W. (1994). Categorization response time with multidimensional stimuli. *Perception & Psychophysics*, **55**, 11-27.

Ashby, F. G., Tein, J.-Y., & Balakrishnan, J. D. (1993). Response time distributions in memory scanning. *Journal of Mathematical Psychology*, **37**, 526-555.

Ashby, F. G., & Townsend, J. T. (1980). Decomposing the reaction time distribution: Pure insertion and selective influence revisited. *Journal of Mathematical Psychology*, **2**, 93-123.

Baddeley, A. D., & Ecob, J. R. (1973). Reaction time and short-term memory: Implications of repetition effects for the high-speed exhaustive scan hypothesis. *Quarterly Journal of Experimental Psychology*, **25**, 229-240.

Balakrishnan, J. D., & Ashby, F. G. (1992). Subitizing: Magical numbers or mere superstition? *Psychological Research*, **54**, 80-90.

Barlow, R. E., & Proschan, F. (1975). *Statistical theory of reliability and life testing: Probability models*. New York: Holt, Rinehart, & Winston.

Bloxom, B. (1984). Estimating response time hazard functions: An exposition and extension. *Journal of Mathematical Psychology*, **28**, 401-420.

Bloxom, B. (1985). A constrained spline estimator of a hazard function. *Psychometrika*, **50**, 301-321.

Burbeck, S. L., & Luce, R. D. (1982). Evidence from auditory simple reaction times for both change and level detectors. *Perception & Psychophysics*, **32**, 117-133.

Carlson, R. A., & Schneider, W. (1989). Acquisition context and the use of causal rules. *Memory & Cognition*, **17**, 240-248.

Colonius, H. (1993). *The instance theory of automaticity: Why the Weibull?* Unpublished manuscript, Institut fur Kognitionsforschung, Oldenburg University.

Dahiya, R. C., & Gurland, J. (1972). Goodness of fit tests for the gamma and exponential distributions. *Technometrics*, **14**, 791-801.

Donders, F. C. (1969). On the speed of mental processes (W. G. Koster, Trans.). In W. G. Koster (Ed.), "Attention and performance II," *Acta Psychologica*, **30**, 412-431. (Original work published 1868)

Dzhafarov, E. N. (1993). Grice-representability of response time distribution families. *Psychometrika*, **58**, 281-314.

Falmagne, J.-C. (1965). Stochastic models for choice reaction time with applications to experimental results. *Journal of Mathematical Psychology*, **2**, 77-124.

Falmagne, J.-C., Cohen, S. P., & Dwivedi, A. (1975). Two-choice reactions as an ordered memory scanning process. In P. M. A. Rabbitt & S. Dornic (Eds.), *Attention and performance V* (pp. 296-344). San Diego, CA: Academic Press.

Fisher, D. L., & Goldstein, W. M. (1983). Stochastic PERT networks as models of cognition: Derivation of the mean, variance, and distribution of reaction time using order-of-processing (OP) diagrams. *Journal of Mathematical Psychology*, **27**, 121-151.

Fisher, R. A., & Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, **24**, 180-190.

Gescheider, G. A., Wright, J. H., Weber, B. J., Kirchner, B. M., & Milligan, E. A. (1969). Reaction time as a function of the intensity and probability of occurrence of vibrotactile signals. *Perception & Psychophysics*, **5**, 18-20.

Gleser, L. J. (1989). The gamma distribution as a mixture of exponential distributions. *American Statistician*, **43**, 115-117.

Green, D. M., & Luce, R. D. (1971). Detection of auditory signals occurring at random times: III. *Perception & Psychophysics*, **9**, 257-268.

Grice, G. R. (1968). Stimulus intensity and response evocation. *Psychological Review*, **75**, 359-373.

Grice, G. R. (1972). Application of a variable criterion model to auditory reaction time as a function of the type of catch trial. *Perception & Psychophysics*, **12**, 103-107.

Heath, R. A., & Wilcox, C. H. (1990). A stochastic model for interkeypress times in a typing task. *Acta Psychologica*, **75**, 13-39.

Heathcote, A., Popiel, S. J., & Mewhort, D. J. K. (1991). Analysis of response time distributions: An example using the Stroop task. *Psychological Bulletin*, **109**, 340-347.

Hockley, W. E. (1984). An analysis of response time distributions in the study of cognitive processes. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **6**, 598-615.

Hohle, R. H. (1965). Inferred components of reaction times as functions of foreperiod duration. *Journal of Experimental Psychology*, **69**, 382-386.

Langston, W., Ohnesorge, C., Kruley, P., & Haase, S. J. (1994). Changes in subject performance during the semester: An empirical investigation. *Psychonomic Bulletin & Review*, **1**, 258-263.

Levinthal, D. A., & Fichman, M. (1988). Dynamics of interorganizational attachments: Auditor-client relationships. *Administrative Science Quarterly*, **33**, 345-369.

Lewinsohn, P. M., Zeiss, A. M., & Duncan, E. M. (1989). Probability of relapse after recovery from an episode of depression. *Journal of Abnormal Psychology*, **98**, 107-116.

Lindsay, B. G., & Roeder, K. (1992). Residual diagnostics for mixture models. *Journal of the American Statistical Association*, **87**, 785-795.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, **95**, 492-527.

Logan, G. D. (1992). Shapes of reaction-time distributions and shapes of learning curves: A test of the Instance Theory of automaticity. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 883-914.

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

Marley, A. A. J., & Colonius, H. (1992). The "horse race" random utility model for choice probabilities and reaction times, and its competing risks interpretation. *Journal of Mathematical Psychology*, **36**, 1-20.

McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review*, **86**, 287-330.

McGill, W. J., & Gibbon, J. (1965). The general gamma distribution and reaction times. *Journal of Mathematical Psychology*, **2**, 1-18.

Miller, J. O. (1988). Discrete and continuous models of information processing: Theoretical distinctions and empirical results. *Acta Psychologica*, **67**, 191-257.

Miller, J. O. (1993). A queue-series model for reaction time, with discrete-stage and continuous-flow models as special cases. *Psychological Review*, **100**, 702-715.

Murdock, B. B. (1985). An analysis of the strength-latency relationship. *Memory & Cognition*, **13**, 511-521.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive*

skills and their acquisition (pp. 1-55). Hillsdale, NJ: Erlbaum.
Ollman, R. (1966). Fast guesses in choice reaction time. *Psychonomic Science*, **6**, 155-156.
Parzen, E. (1962). On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics*, **33**, 1065-1076.
Pavel, M. (1990, November). *A statistical model of preattentive visual search*. Paper presented at the 31st Annual Meeting of the Psychonomic Society, New Orleans.
Posner, M. I. (1982). Cumulative development of attentional theory. *American Psychologist*, **37**, 168-179.
Proschan, F. (1963). Theoretical explanation of observed decreasing failure rate. *Technometrics*, **5**, 375-383.
Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, **85**, 59-108.
Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, **86**, 446-461.
Ratcliff, R. (1980). A note on modeling accumulation of information when the rate of accumulation changes over time. *Journal of Mathematical Psychology*, **21**, 178-184.
Ratcliff, R. (1988). A note on the mimicking of additive reaction time models. *Journal of Mathematical Psychology*, **32**, 192-280.
Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, **114**, 510-532.
Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review*, **83**, 190-214.
Revelle, W. (1993). Individual differences in personality and motivation: "Non-cognitive" determinants of cognitive performance. In A. Baddeley & L. Weiskrantz (Eds.), *Attention: Selection, awareness and control: A tribute to Donald Broadbent* (pp. 346-373). Oxford: Oxford University Press.
Roberts, S., & Sternberg, S. (1994). The meaning of additive reaction-time effects: Tests of three alternatives. In S. Kornblum & D. E. Meyer (Eds.), *Attention and performance XIV* (pp. 611-653). Cambridge, MA: MIT Press.
Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, **84**, 1-66.
Schweickert, R. (1978). A critical path generalization of the additive factor method. *Journal of Mathematical Psychology*, **18**, 105-139.
Schweickert, R. (1980). Critical-path scheduling of mental processes in a dual task. *Science*, **209**, 704-706.
Schweickert, R. (1983). Latent network theory: Scheduling of processes in sentence verification and the Stroop effect. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **9**, 353-383.
Schweickert, R., & Townsend, J. T. (1989). A trichotomy: Interactions of factors prolonging sequential and concurrent mental processes in stochastic discrete mental (PERT) networks. *Journal of Mathematical Psychology*, **33**, 328-347.
Smith, P. L., & Vickers, D. (1988). The accumulator model of two-choice discrimination. *Journal of Mathematical Psychology*, **32**, 135-168.
Sternberg, S. (1966). High-speed scanning in human memory. *Science*, **153**, 652-654.
Sternberg, S. (1967). Two operations in character recognition: Some evidence from reaction-time measurements. *Perception & Psychophysics*, **2**, 45-53.
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.
St. James, J. D., & Eriksen, C. W. (1991). Response competition produces a "fast same effect" in same–different judgments. In G. R. Lockhead & J. R. Pomerantz (Eds.), *The perception of structure: Essays in honor of Wendell R. Garner* (pp. 157-168). Washington, DC: American Psychological Association.
Strayer, D. L., & Kramer, A. F. (1990). An analysis of memory-based theories of automaticity. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 291-304.
Tallis, G. M. (1969). The identifiability of mixtures of distributions. *Journal of Applied Probability*, **6**, 389-398.
Townsend, J. T. (1972). Some results concerning the identifiability of parallel and serial processes. *British Journal of Mathematical Psychology*, **25**, 168-199.
Townsend, J. T. (1974). Issues and models concerning the processing of a finite number of inputs. In B. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition* (pp. 133-168). Hillsdale, NJ: Erlbaum.
Townsend, J. T. (1990a). 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. (1990b). Truth and consequences of ordinal differences in statistical distributions: Toward a theory of hierarchical inference. *Psychological Bulletin*, **108**, 551-567.
Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes*. New York: Cambridge University Press.
Townsend, J. T., & Roos, R. N. (1973). Search reaction time for single targets in multiletter stimuli with brief visual displays. *Memory & Cognition*, **1**, 319-332.
Townsend, J. T., & Schweickert, R. (1989). Toward the trichotomy method of reaction times: Laying the foundation of stochastic mental networks. *Journal of Mathematical Psychology*, **33**, 309-327.
Townsend, J. T., & Thomas, R. D. (1994). Stochastic dependencies in parallel and serial models: Effects on systems factorial interactions. *Journal of Mathematical Psychology*, **38**, 1-34.
Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, **12**, 97-136.
Treisman, A., Vieira, A., & Hayes, A. (1992). Automaticity and preattentive processing. *American Journal of Psychology*, **105**, 341-362.
Ulrich, R., & Miller, J. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology*, **123**, 34-80.
Van Zandt, T., Colonius, H., & Proctor, R. W. (1995). *A Poisson race model of "same"–"different" matching*. Manuscript submitted for publication.
Vickers, D. (1979). *Decision processes in visual perception*. New York/London: Academic Press.
Vickers, D., Caudrey, D., & Willson, R. J. (1971). Discriminating between the frequency of occurrence of two alternative events. *Acta Psychologica*, **35**, 151-172.
Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, **1**, 202-238.
Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception & Performance*, **15**, 419-433.
Woodworth, R. S. (1938). *Experimental psychology*. New York: Holt.
Yantis, S., Meyer, D. E., & Smith, J. E. K. (1991). Analyses of multinomial mixture distributions: New tests for stochastic models of cognition and action. *Psychological Bulletin*, **110**, 350-374.
Yellott, J. I. (1967). Correction for guessing in choice reaction time. *Psychonomic Science*, **8**, 321-322.
Yellott, J. I. (1971). Correction for fast guessing and the speed-accuracy trade-off in choice reaction time. *Journal of Mathematical Psychology*, **8**, 159-199.

NOTES

1. Although the diffusion model assumes that the drift rate changes during the course of a trial (Ratcliff, 1980), for the purposes of this discussion we will assume that the drift rate is selected at the beginning of a trial and fixed and that the distribution of drift rates occurs over trials.

2. One of the reviewers pointed out that our choice of a scenario makes it appear that mixtures arise only on an artifactual basis. This was not our intent. An equivalent scenario arises in the case where two stimuli are presented in a choice RT task, and sequential effects are observed such that the speed of responding depends both on the stimulus presented and on the accuracy of the previous response. Collapsing the data across these conditions produces a mixture. For instance, in the long display duration, the fastest RTs would arise when the previous response was correct and the previous stimulus was the same as the current one. The slowest RTs would result when an error was made on the previous trial. In the short display durations, errors would increase in frequency, shifting the mixture from faster to slower RT distributions. Our "researcher" scenario makes for more entertaining reading, and we hope the reader will indulge us. The potential trivial-

ity of the mixture produced by careless programming should not be taken as triviality of the mixture problem.

3. It does, however, suggest that another way to test for a single exponentially inserted stage might be to simply add exponential deviates (drawn from a distribution with a mean equal to the mean increase in RTs observed in the two experimental conditions) to the fast RTs collected in the first condition. Nonparametric tests could then be applied to the slow RTs collected in the second condition and the fast RTs plus the exponential deviate to determine if the two data sets were sampled from the same distribution.

4. Note that, although we have used the gamma distribution to represent multiple-process models early in the discussion, it represents a single process in the present model. This is because the number of exponentials that compose it does not change across experimental conditions but stays the same. In this case, the shape parameter is representing a threshold that does not vary with the amount of practice. If the number of exponential "stages" decreased with practice, the gamma would represent a multiple-process IES model.

APPENDIX A
Mean RT Predictions and the Existence of Single-Process Mimics

There are an infinite number of models that can predict a given pattern of mean RTs in the absence of any other constraints.

Suppose that the central process under scrutiny produces finishing times T that are distributed as $G(t|\Lambda)$. The parameter Λ is also a random variable with some density function $h(\lambda|\beta)$ defined over $S \subseteq \Re$. Then for some fixed β, an experimenter observes a continuous mixture of finishing times distributed as $F(t|\beta)$, where

$$F(t|\beta) = \int_S G(t|\lambda)\, h(\lambda|\beta)\, d\lambda.$$

For the purposes of this discussion, we call G the base distribution of F and h the parameter density of Λ. In general, G is specified by an explicit model of the process. If

$$\int_0^\infty t\, dG\,(t|\lambda) < \infty,$$

then $E_G[T|\lambda]$ is defined. Also, if $E_G[T|\lambda]$ is a bounded function of λ and the mean of Λ is defined, then the mean finishing time is also defined as

$$E_F[T|\beta] = \int_S E_G[T|\lambda] h(\lambda|\beta)\, d\lambda,$$

a function of β.

If $E_F[T|\beta]$ is continuous on $[a, b]$ and we observe a set of mean RTs $\{\mu_1, \mu_2, \ldots, \mu_n\}$ such that all $\mu_i \in [E_F[T|a], E_F[T|b]]$, then there exists a set of points $\{\beta_1, \beta_2, \ldots, \beta_n\}$ contained in $[a, b]$, not necessarily unique, such that $\mu_1 = E_F[T|\beta_1]$, $\mu_2 = E_F[T|\beta_2]$, $\ldots$, $\mu_n = E_F[T|\beta_n]$. This is the intermediate value theorem from elementary calculus. Notice that for a specific model G and a specific parameter variability problem h, there may be many sets of parameters $\{\beta_1, \beta_2, \ldots, \beta_n\}$ that will produce $\{\mu_1, \mu_2, \ldots, \mu_n\}$. This is true, for example, if $E_F[T|\beta]$ is nonmonotonic on $[a, b]$. Therefore, there exists a problem of indeterminacy at the level of the mean RTs even when the model and parameter variability are completely specified.

Now consider an entirely different model G', with (perhaps) a completely different parameter density h'. If the expected finishing time for this model is defined and is a bounded function of the varying parameter, and, likewise, the mean of this parameter is defined, then the observed finishing times have expected value $E_{F'}[T|\beta']$. All that is required for this completely different model to perfectly mimic the mean RT predictions of the previous model is for $E_{F'}[T|\beta']$ to be continuous on $[a, b]$ and $\mu_i \in [E_{F'}[T|a], E_{F'}[T|b]]$, $i = 1, 2, \ldots, n$. Applying the intermediate value theorem again assures us that there exists at least one set of points $\{\beta'_1, \beta'_2, \ldots, \beta'_n\}$ contained in $[a, b]$ that will produce $\{\mu_1, \mu_2, \ldots, \mu_n\}$.

As a simple example, suppose that $G[t|\Lambda) = \Phi(t-\Lambda)/\sigma]$, the normal distribution with mean Λ and standard deviation σ. The mean Λ varies exponentially with rate β, where $\beta > 0$. The resulting mixture distribution $F(t|\beta)$ is then an ex-Gaussian with mean $E_F[T|\beta] = 1/\beta$. The function $E_F[T|\beta]$ is continuous on $(0, \infty)$, and, a fortiori, any observed mean RT μ_i must be contained in the interval $(E_F[T|\infty], E_F[T|0])$. Therefore, this model can produce any pattern of mean RTs $\{\mu_1, \mu_2, \ldots, \mu_n\}$ by setting $\beta_1 = 1/\mu_1$, $\beta_2 = 1/\mu_2$, and so on. Now notice that the mean RTs can also be produced by a completely different model, where $F'(t|\beta') = 1 - e^{-t\beta'}$ and it is assumed that the rate parameter β' is constant. Because $E_F[T|\beta] = E_{F'}[T|\beta']$, setting $\beta_1 = \beta'_1$, $\beta_2 = \beta'_2$, and so on allows no way to discriminate (at the level of the mean RTs) between the model that presumes normally distributed finishing times and an exponentially varying parameter and the model that presumes exponentially distributed finishing times and no parameter variability.

APPENDIX B
Power Function Decrease of the Mean, Standard Deviation, and Quantiles of a Variable

Weibull distributed finishing times are not a necessary assumption to observe power function decreases in mean, standard deviation, and quantiles of the RTs. Suppose that a finishing time random variable T is distributed as $F(t)$, $E[T] = \mu$, and the standard deviation of T is σ. Now consider the random variable $T_N = N^{-c}\, T$. It can easily be seen that $E[T_N] = N^{-c}\mu$, the standard deviation of T_N is $N^{-c}\sigma$, and, after performing the change of variables, the distribution function of T_N is $F(N^c t)$. The distribution function F therefore generates a scale parameter family of distributions, and the distributions $F(N^c t)$, $N = 1, 2, \ldots$, are contained in that family. Regardless of the original distribution F, the means, standard deviations, and quantiles of T_N, $N = 1, 2, \ldots$ all decrease as a power function of n with the same exponent c.

APPENDIX C
Derivation of the Single-Process Model of Automaticity

The accumulator mechanism that drives the response-selection process receives units of information at exponentially distributed interarrival times. The rate at which units are received is Λ, a random variable that follows a gamma distribution with shape parameter j and rate parameter $\gamma(N)$. The function $\gamma(N)$ is a positive function of $N = 1, 2, \ldots$ defined over $(0,\infty)$. A response is initiated after k units are recorded, which defines the base distribution of the finishing times also as a gamma, with shape parameter k and rate parameter Λ. From Equation 1, the observed RTs are then distributed as

$$F_N(t|\gamma(N)) = \int_0^\infty G\,(t|\lambda) h[\lambda|\gamma(N)]\, d\lambda,$$

or working from the densities,

$$f_N(t|\gamma(N)) = \int_0^\infty g(t|\lambda)h[\lambda|\gamma(N)]d\lambda$$

$$= \int_0^\infty \frac{\lambda(\lambda t)^{k-1}e^{-\lambda t}}{\Gamma(k)} \cdot \frac{\gamma(N)[\gamma(N)\lambda]^{j-1}e^{-\gamma(N)\lambda}}{\Gamma(j)} d\lambda$$

$$= \frac{\Gamma(k+j)}{\Gamma(k)\Gamma(j)} \cdot \frac{1}{\gamma(N)} \left[\frac{t}{t+\gamma(N)}\right]^{k-1} \left[\frac{\gamma(N)}{t+\gamma(N)}\right]^{j+1}.$$

Notice that the transformation $T_N/(T_N + \gamma(N))$ gives a beta random variable with parameters k and j.

To find the m^{th} moment of T_N, multiply the above expression by t^m and make the change of variable $u = t/[t + \gamma(N)]$. Integrating over u gives

$$E[T_N{}^m] = \gamma(N)^m \frac{\Gamma(k+m)\Gamma(j-m)}{\Gamma(k)\Gamma(j)}.$$

Therefore, the moments of T_N are not defined beyond $j-1$. Also,

$$E[T_N] = \gamma(N)\frac{k}{j-1}$$

and

$$\text{Var}[T_N] = \gamma(N)^2 \frac{(k+1)k}{(j-1)(j-2)} - \gamma(N)^2 \frac{k^2}{(j-1)^2}$$

$$= \gamma(N)^2 \frac{k(k+j-1)}{(j-1)(j-2)}.$$

Define $\gamma(N)$ as a power function of N. Thus, the mean and standard deviation of T_N decrease as a power function of N. Inserting $\gamma(N) = \gamma(1)N^{-c}$ into the equation for the density, it can be seen that the power function decrease for the quantiles must hold as well.

APPENDIX D
Deriving Predictions for the Single-Process Visual Search Model

For stimuli composed of conjunctions of features, such as color, size, or shape, each feature value (e.g., red) appearing in the target is assumed to contribute a normally distributed amount of activation (with mean u and standard deviation s) to the locations in the attention map corresponding to the display locations where they are found. These activations are independent across display locations and stimulus dimensions, resulting in an overall mean activation level and activation variance that are equal to the sums of the individual means and variances of the activation distribution. So, in the simple conjunction case, each nontarget location in the attention map is given an average activation u with standard deviation s. When no target is present, the per-location average is then u and the per-location standard deviation (standard error) is s. When a single conjunction target is present, one of the map locations contains on average a level $2u$ of activation with variance $2s^2$. A triple conjunction target location has mean activation $3u$ and variance $3s^2$. The per location mean activation level across all locations in the attention map is therefore u for negative displays, $(N+1)u/N$ for simple conjunction positive displays, and $(N+2)u/N$ for triple conjunction positive displays. For the triple conjunction condition in which the distractors share two target features, the mean activation levels are $2u$ and $(2N + 1)u/N$ for the negative and positive cases, respectively.

To estimate the per-location activation level variance, we used the mean of the individual location variances. We also could have used the standard error of the activation mean by dividing each of the mean variances by N. There was no particular reason that we chose the mean estimate over the standard error, except perhaps the larger variances allowed a greater range of accuracy across N. The standard errors worked equally well, producing approximately equal r^2 and *rmse* measures for the fits to the RTs as did the fits using the mean variances. The mean variances were equal in proportion to the mean activation levels; for instance, the triple conjunction (one feature shared) target-present distribution had variance $(N+2)s^2/N$.

An unbiased criterion is set at the point where the probability of correctly detecting a target is equal to the probability of correctly determining that no target is present. Equating the z scores for the target-present and target-absent activation densities yields the following formula for the criterion:

$$c = \frac{s_-\mu_+ + s_+\mu_-}{s_- + s_+},$$

where s_j and m_j $(j = +,-)$ indicate the standard deviations and means estimated for the target-present and target-absent activation densities in each condition.

The $s_j s$, $\mu_j s$, and criteria were all estimated in this way, setting u equal to one and using a least squares minimization to estimate s and the remaining parameters of the mean RT function.

We calculated the mean RTs as follows. First, we assumed that processing time decreases as a function of the distance of the activation level a from the criterion c:

$$E[RT_j \mid a,c] = \gamma e^{-\lambda_j|a-c|} + b,$$

where the subscript j indicates the asymmetry between positive and negative responses. This function gives a reasonable approximation of the decrease in mean RT as stimulus intensity falls farther and farther from criterion (e.g., Ashby et al., 1994; Gescheider et al., 1969). The decay rate λ is assumed to be greater for positive responses, capturing the faster RTs often used as evidence for a self-terminating search. One rationale for such an assumption is the increased variance of the positive displays: because more target activation levels fall in the nontarget region than vice versa, an activation level sampled below criterion is more likely to be in error, requiring more caution and hence slower RTs.

The sampled activation level a is a normally distributed random variable with mean μ_+ or μ_-, and standard deviation σ_+ or σ_-, conditioned on whether a target is present (positive) or absent (negative) in the display. Averaging across the variable activation level a, the mean correct RTs predicted by the mixture are

$$E[RT_-|c] = \int_{-\infty}^{\infty} (\gamma e^{-\lambda_-|a-c|} + b)\left[(1/\sqrt{2\pi}\sigma_-)e^{-(a-c)^2/2\sigma_-^2}\right]da$$

$$= \frac{\gamma e^{[(\lambda_-^2\sigma_-^2)/2]-\lambda_-(c-\mu_-)}\Phi[(c-\mu_- - \lambda_-\sigma_-^2)/\sigma_-]}{\Phi[(c-\mu_-)/\sigma_-]} + b,$$

for a negative response, where Φ is the standard normal distribution function, and similarly

$$E[RT_+ \mid c] = \frac{\gamma e^{[(\lambda_+^2\sigma_+^2)/2]-\lambda_+(\mu_+-c)}\Phi[(\mu_+ - c - \lambda_+\sigma_+^2)/\sigma_+]}{\Phi[(\mu_+ - c)/\sigma_+]} + b$$

for a positive response. Notice that this model not only predicts the mean RTs but also the accuracies via the terms $\Phi[(c-\mu_-)/\sigma_-]$ and $\Phi[(\mu_+-c)/\sigma_+]$.

A total of six free parameters were estimated. These parameters generated predictions for 18 mean RTs and 18 accuracies, leaving 30 degrees of freedom.

APPENDIX E
Scale Parameter Families and the Insertion of a Serial Subprocess

The existence of a number of subprocesses within a larger cognitive task implies the existence of a sequence of random variables that represent the duration of each subprocess. As the task demands are increased, subprocess durations are increased, or additional subprocesses are inserted calling for additional random variables to represent them. The case where additional subprocesses are inserted is considered a multiprocess model.

Townsend and Schweickert (1989, Theorem 1) have shown that incrementing the duration of a process by adding another positive random variable to it (the multiprocess model) is logically equivalent to imposing an order on the distribution functions representing the pre- and postincremented finishing-time distributions. Thus, any sequence of ordered RT distributions could have been produced by a systematic addition of new (dependent) subprocesses. For example, in a memory search task, the addition of a distractor might introduce an additional comparison subprocess, adding a new random variable to the finishing time, which in turn orders the RT distributions.

Note, however, that any scale parameter family of distributions will produce ordered distributions if the experimental manipulation (incrementing procedure) causes a systematic increase in the scale parameter. Let the random variable X be distributed as F, and let σX be distributed as F_σ. Thus, $\sigma > 0$ is a scale parameter, and F then generates the family of distributions $\mathbf{F}_s = \{F_\sigma: \sigma > 0\}$. Choose two scale values σ and τ such that $\sigma < \tau$. These values define two members of $\boldsymbol{F}_s$, which can be written in terms of each other:

$$F_\sigma(x) = F_\tau(\tau x/\sigma).$$

That this is true can be seen by defining two random variables Y and Z such that $Y = \sigma X$ and $Z = \tau X$, which implies that $Y = (\sigma/\tau)Z$. Thus, $(\sigma/\tau)Z$ must be distributed as F_σ, which, by a change of variables, must equal $F_\tau(\tau x/\sigma)$. If an experimental manipulation has the effect of increasing the scale parameter, say, from σ to τ, the distribution functions of the finishing times will be ordered, since $F_\tau(x) < F_\tau(\tau x/\sigma)$ by the monotonic nature of the distribution function F_τ. By Townsend and Schweickert's result, it also implies the existence of another positive random variable $U = (\tau - \sigma)X$ that, when added to Y, gives the variable Z. (Note that U is dependent on both Y and Z.) There are then two equivalent representations of the effect of the experimental manipulation. The first one is where the manipulation lengthened the scale parameter of the process; the second one is where a new subprocess of duration $U = (\tau - \sigma)X$ was added to the old process duration Y.

Thus, any single-subprocess model that orders the distribution functions from one experimental condition to the next by increasing the rate parameter of the process can be represented by a (dependent) multiprocess model in which the experimental manipulations have the effect of increasing the number of subprocesses.

APPENDIX F
The Existence of a Mixture Representation for any Density Function

Suppose that a given set of RTs has density function $f(t)$. Does $f(t)$ then give any information about the process underlying the generation of those RTs? If for any density $f(t)$, we can find an arbitrary mixture representation, given as

$$f(t) = \int_{\alpha \in S} g(t|\alpha)h(\alpha)d\alpha,$$

for some densities g and h, where α takes on values in some finite interval $S \subseteq \Re$, then the answer is no. It might be obvious to some readers that, indeed, for a particular $f(t)$, there are an infinite number of mixture representations. This follows from the fact that any function can be expanded as a general Fourier series using an orthonormal set of basis functions. However, it is not that obvious that there exist density functions (i.e., positive and of unit mass) g and h that will produce f. In other words, the expansion of f may not necessarily be based on random variables that could represent finishing time distributions.

Consider the following. For the given function $f(t)$, select an arbitrary parameter α taking on values in the finite interval $[a,b]$. (If $[a,b]$ is infinite for some parameter of a particular distribution, then a suitable transformation of the parameter can be made.) Now we must construct an arbitrary function $h(\alpha)$ that satisfies the following three conditions:

1. $h(\alpha) > 0$ for all $\alpha \in [a,b]$.