Psychological Review
1999, Vol. 106, No. 2, 261-300

0033-295X/99/$3.00

Connectionist and Diffusion Models of Reaction Time

Roger Ratcliff
Northwestern University

Trisha Van Zandt
Johns Hopkins University

Gail McKoon
Northwestern University

Two connectionist frameworks, GRAIN (J. L. McClelland, 1993) and brain-state-in-a-box (J. A. Anderson, 1991), and R. Ratcliff's (1978) diffusion model were evaluated using data from a signal detection task. Dependent variables included response probabilities, reaction times for correct and error responses, and shapes of reaction-time distributions. The diffusion model accounted for all aspects of the data, including error reaction times that had previously been a problem for all response-time models. The connectionist models accounted for many aspects of the data adequately, but each failed to a greater or lesser degree in important ways except for one model that was similar to the diffusion model. The findings advance the development of the diffusion model and show that the long tradition of reaction-time research and theory is a fertile domain for development and testing of connectionist assumptions about how decisions are generated over time.

Research aimed at investigating how information is processed over time has had a long and influential history in psychology. In 1938 in his general textbook, Woodworth discussed simple and choice reaction time, the behaviors and shapes of reaction-time distributions, individual differences in reaction time, and the effects on reaction time of experimental variables such as stimulus intensity. Several of these topics are raised again in this article. In the 1960s, when the cognitive revolution gave rise to modern cognitive psychology, reaction time entered the spotlight as a major dependent variable. Since then, considerable effort has been devoted to the development of theories to explain how information is processed over time to yield mean response times, distributions of response times, and accuracy levels. Current theoretical issues include, for example, serial versus parallel processes and continuous versus discrete processes, and efforts continue toward comprehensive theories of the time course of processing. A summary of the state of reaction-time theory is presented in Luce (1986). Perhaps the main difficulty in recent modeling has been that two dependent variables, reaction time and the probability of correct versus error responses, have to be modeled in the same, integrated framework.

Connectionist models are a relatively new class of models and a surge in development and testing of them has taken place in the last 10 years. These models offer the promise of explanations of how cognitive tasks are learned. For most of the models, learning is the result of many individual trials with stimuli, each trial with feedback about whether the model's response was correct. The processes by which the response to a stimulus is chosen are usually assumed to be parallel, interactive, nonlinear, and continuous. These processing characteristics are theoretical choices that have been well examined in reaction-time modeling. Therefore, it is potentially fruitful for connectionist models to meet reaction-time models in a joint effort at theory development and competitive model testing and evaluation.

Carrying reaction-time research forward to meet the relatively new domain of connectionist modeling was one purpose of the investigations described in this article. Specifically, we asked whether connectionist models could accommodate the wide-ranging kinds of data that have been critical in the reaction-time domain and at the same time account for learning. A second purpose was to test and further develop a more standard model, Ratcliff's (1978) diffusion model. Standard information-processing models and connectionist models have different insights to offer, and the fullest advantage of these insights can be gained when both kinds of models are pushed as far as they possibly can be. This goal can best be accomplished in an arena of investigation that allows simultaneous testing of both kinds of models. For the research described in this article, the arena we chose was a simple signal detection paradigm.

Connectionist models assume that the decision required on each trial of an experimental task comes about by processes that integrate and accumulate information over time. For early connectionist models, an Achilles' heel was their failure to match this assumption to specific mechanisms that could predict a full range of empirical measures of the time course of processing, including

Roger Ratcliff and Gail McKoon, Psychology Department, Northwestern University; Trisha Van Zandt, Psychology Department, Johns Hopkins University.

This research was supported by National Institute of Mental Health Grants HD MH44640 and HD MH00871, National Institute for Deafness and Other Communication Disorders Grant R01-DC01240, and National Science Foundation Grants SBR9221940 and SBR9702291. We wish to thank James McClelland for many useful suggestions and discussions of this research.

Correspondence concerning this article should be addressed to Roger Ratcliff, Psychology Department, Northwestern University, Evanston, Illinois 60208.

the probabilities of one response versus another, response times, the interactions among error versus correct response times, and the shapes of response-time distributions. The models were successful at combining structural assumptions about how information is represented with algorithms to learn a task, producing outputs that qualitatively matched either the mean accuracy of responses or mean reaction time, but they were not originally designed to provide a simultaneous account of a more complete range of data.

In this article, we focus on two more recent connectionist frameworks that have been designed specifically to deal with the full range of measures. One, Anderson's (1991) brain-state-in-a-box model (BSB), is an autoassociative matrix model. The other, McClelland's (1993) GRAIN framework, provides a list of principles with which to begin exploration of processing time. Using the principles, we constructed two multilayer models for examination. Both the BSB model and the GRAIN models are based on an iterative algorithm that takes a variable number of steps to reach a decision such that a stimulus input does not always produce the same output or take the same amount of time. These features potentially allow the models both to learn a cognitive task and to predict the multiple aspects of performance that are reflected in the shapes of distributions of response times and speed and accuracy interactions. This is an important advance. The attempt to deal with the complete range of reaction-time data in the context of a framework originally developed to examine learning adds a new degree of complexity to connectionist models. Most often, connectionist models in psychology have attempted to explain the behavior of only one dependent variable, usually a variable closely related to learning such as probability of a correct response.

There are a number of traditional, nonconnectionist, information-processing models that have been shown to provide a good account of the time course of processing (see, e.g., Luce, 1986; Townsend & Ashby, 1983). These models do not provide accounts of learning; their strength is that they deal with the multiple dependent variables that can be used to measure decision processes. The diffusion model (which can be seen as an extension of earlier random walk models; Laming, 1968; Link & Heath, 1975; Stone, 1960) was developed by Ratcliff (1978, 1980, 1981, 1988) for two-choice tasks. It was chosen for comparison to connectionist models for three reasons. First, it is a member of the general class of random walk models that provide better accounts of many experimental results than counter models with absolute criteria or various serial or parallel models (see Luce, 1986). Second, the diffusion model has been successfully applied to a wide range of experimental paradigms, accurately accounting for mean reaction times, error rates, the shapes of reaction-time distributions, and the effects of several deadline and response signal manipulations. Other exemplars of the class of random walk models (Laming, 1968; Link & Heath, 1975; Stone, 1960) can also fit many aspects of data well but do not have the explicit mathematical expressions for describing characteristics of the data that the diffusion model does. For example, for the diffusion model, there are explicit formulas for the distributions of reaction times and for the distributions of unterminated processes at any point in time (Meyer, Irwin, Osman, & Kounios, 1988; Ratcliff, 1988). Third, the diffusion model was designed to explain fast, single step, as opposed to multistep, decision processes, and in this respect it is similar to the connectionist models to be evaluated.

The assumption underlying the diffusion model is that information is accumulated continuously over time. There are two boundaries on the accumulation process, one for each of the two possible response choices. Information is accumulated from a starting point toward either of the boundaries and when sufficient information is accumulated that one of the boundaries is crossed, a decision is made. Noise in the process causes variations in the rate at which information is accumulated over time, so that the same stimulus presented on different occasions does not always lead to the same decision or require the same amount of time for a decision. These aspects of the model allow it to account for the shapes of distributions of response times and speed–accuracy interactions.

We chose a simple signal detection task to provide a comprehensive data base to compare, contrast, and test the diffusion model and the models from the connectionist frameworks. We had three simultaneous goals: first, to test connectionist assumptions about the time course of the decision processes on individual trials and to test these assumptions in the same empirical context as assumptions about learning were tested; second, to extend the diffusion model to a new experimental task, evaluating its ability to deal with correct and error reaction times as well as reaction-time distributions and response probabilities; and third, to compare the connectionist models and the diffusion model, in an empirical situation in which strong predictions could be made about data. We begin by describing the experimental task we chose and then present data from it. Then the diffusion model, the two connectionist models derived from the GRAIN framework, and the BSB model are presented and applied to the data.

The Signal Detection Paradigm

An important choice for evaluation and comparison of connectionist and standard models of reaction time is the experimental paradigm that will provide the testing ground. To choose the experimental task for our research, six main criteria were adopted. First, the task had to involve relatively simple stimuli so that assumptions about how the stimuli were represented would not interfere with examinations of reaction time and accuracy. Second, the task had to allow examination of the full range of measures against which reaction-time models are routinely tested: reaction times for both correct and incorrect responses, covariations of accuracy and reaction time, and the shapes of the distributions of reaction times and their hazard functions. Third, response probability or accuracy had to span the range from near chance to near ceiling so that the full range of correct and error reaction times could be examined as a function of response probability. Fourth, the task had to be representative of a wide range of experimental tasks, and it had to produce data with typical variations in and interactions among all the standard speed and accuracy measures. Fifth, the task had to have a learning component by which statistical properties of the stimuli might be expected to engage the learning mechanisms of the connectionist models, and sixth, the learning component had to allow examination of the sequential effects across learning trials that would be predicted by connectionist models.

The task we chose met all of the criteria just listed. It is a signal detection paradigm adapted from Espinoza-Varas and Watson (1994; see also precursors, e.g., Lee & Janke, 1964; Smith & Vickers, 1988; Vickers, 1979; Vickers, Caudrey, & Willson, 1971, and probability learning paradigms, Atkinson, Bower, & Crothers, 1966; Estes, 1957, 1964). (It also turns out that the paradigm is the 1-D analog of the general recognition randomization technique used by Ashby & Gott, 1988.) In our application, on each trial, an array of asterisks was presented on a computer screen and an observer was asked to decide whether the number of asterisks presented in the display was "high" or "low." The number of asterisks that was presented was chosen from one of two distributions of numbers, a high distribution and a low distribution, each distribution with fixed mean and standard deviation, and all numbers between 0 and 99. Feedback was given after each trial to tell the observer whether his or her response had correctly indicated the distribution from which the stimulus had been chosen. Other than this feedback, the observer had no information about the distributions. The distributions overlapped substantially, so that even after many trials of feedback, the observer could not be highly accurate. A display of 50 asterisks, for example, might have come from the high distribution on one trial and the low distribution on another.

The signal detection paradigm has many advantages over other possible choices. One is that the variable underlying performance (number of asterisks) can be varied in small steps from a high probability of one response to a high probability of the other response. For example, a display of only 5 asterisks strongly supports the "low" choice, a display of 90 asterisks strongly supports the "high" choice, and a display of 50 asterisks is in the middle, strongly supporting neither choice. In lexical decision, for example, comparable variation is not possible; a stimulus is either a word or it is not. The paradigm offers a great deal of experimental control, while allowing measurement of all the speed and accuracy interactions needed for stringent tests of models.

Another important advantage of the signal detection task is the generality of the results that are obtained with it. First, it provides data typical of a large class of signal detection procedures. For example, Espinoza-Varas and Watson (1994) used two-digit numbers and tones of differing frequency; Lee and Janke (1964) used two-digit numbers, gray scale stimuli, and line lengths; and Vickers (1979) used binary dot patterns, line lengths, lamps flashing at varying rates, and randomly oriented line segments. Ratcliff and Rouder (1998) used patches of black and white pixels varying in brightness, patches of red and green pixels varying from red to green, and two patches of black and white pixels that required same–different judgments of brightness.

The signal detection task is also representative of many cognitive paradigms such as lexical decision, matching tasks, recognition memory, and semantic verification if it is assumed that the basis for decisions in these tasks is a unidimensional continuum. For example, for lexical decision, a model might claim that a "word" decision depends on the amount of activation or familiarity in lexical memory evoked by the presentation of a stimulus and that the response is based on this value (high for a "word" response, low for a "nonword" response). Similarly, in a same–different matching task, a model might claim that decisions are based on the number of perceived differences between two stimulus elements (a large number for a "different" response and a small number for a "same" response; cf. Krueger, 1978). The data from the asterisk signal detection task provides the same speed–accuracy interactions and the same shapes of response-time distributions as are found with other cognitive paradigms. If either the diffusion model or the connectionist models fail to account for some pattern of reaction time or accuracy data from the signal detection task, then they would probably also fail to account for similar patterns from other cognitive tasks.

The signal detection task also allows investigation of what has been a particularly difficult problem for standard reaction-time models. Currently, no reaction-time model has been able to account for the different relations that are observed empirically between correct and error response times. Generally, when accuracy is stressed as being more important than speed and the task is difficult (e.g., difficult perceptual discriminations, Swensson, 1972, or difficult recognition memory tasks, Ratcliff & Murdock, 1976), reaction times for errors are slower than reaction times for correct responses. In contrast, when speed is stressed over accuracy and the task is easy (e.g., choice reaction time, Laming, 1968; Swensson, 1972), error reaction times are faster than correct reaction times (see Luce, 1986, p. 233, for discussion of the available data). In the signal detection task, a task intermediate in difficulty, individual subjects can adopt different criteria. In Experiment 1 below, one subject showed errors faster than correct responses, one showed errors slower than correct responses, and the other two subjects showed a mixture of the two patterns (see also Smith & Vickers, 1988). Thus, the signal detection paradigm offers data suitable for examining what no model has yet successfully explained—the varying relationships between correct and error reaction times.

Overview and Preview

There were four steps in our research. First, in order to get stable results for modeling of the response time and accuracy measures in the signal detection paradigm, we collected a large amount of data from individual subjects (Experiment 1). Second, the diffusion model was fit to the data. The model provided a good account of the data (including accurate quantitative fits) and so served as a demonstration of the tractability of the paradigm for modeling and provided a level of explanation of the data against which to compare the connectionist models. Third, the two models in the GRAIN framework and the BSB model were tested against the data.

In providing a good account of the data, the diffusion model found an invariance across subjects' performance that had not been anticipated. Specifically, it appeared that all subjects based their decisions on the probability that a stimulus number of asterisks was chosen from the high versus the low distribution. This finding, that decisions were based on stimulus probability, became apparent only through application of the diffusion model; it was not directly apparent in the data. For this reason, we added a fourth step to the research: In Experiment 2, we varied the probabilities that the stimuli were chosen from the high versus the low distributions, switching from one set of probabilities to another in the course of a single experimental session. If subjects were indeed basing their decisions on stimulus probability, as the diffusion model indicated for Experiment 1, then their decision processes should appropriately follow the switches in probabilities.

Over the two experiments and all the different measures, there is a large set of tests of the connectionist models and the diffusion model. Before we begin detailed description of the experiments, data, and model applications, we give in the next paragraphs a brief preview of how the models fared.

The Diffusion Model

An unanticipated experimental finding was that subjects provided more radically different patterns of data than expected, both at the broad levels of mean reaction time and accuracy and also in more complex interactions between the reaction time and accuracy measures. The diffusion model provided an explanation of how these different patterns of behavior came about, generating predictions that accurately matched all aspects of each subject's individual data. Moreover, the diffusion model explanation showed how the single underlying variable, the probability that a stimulus had been chosen from the high versus the low distributions, could govern all of the subjects' performance, and the model showed how the large differences in performance could arise even when all of the subjects were basing their responses on this same underlying variable.

In accounting for all of the response-time and accuracy data, the diffusion model was able to explain how different patterns of correct versus error reaction times came about for different subjects, the problem that had not been solved by previous reaction-time models. It is important to find that a model can do this, but more important is the way the diffusion model does it—by assuming that the value of a parameter in the model is variable across trials, not fixed (e.g., Laming, 1968; Ratcliff, 1978, 1981). In other words, the assumption is that some aspects of the decision process—for example, the starting point or the average rate at which the accumulation of information approaches a decision boundary—are not constant across trials. The finding that a model can gain significant power to handle data with variable rather than fixed parameter values offers an avenue for modeling not previously well exploited in cognitive theory (see Van Zandt & Ratcliff, 1995).

The BSB Model

The BSB model was reasonably successful in dealing with mean reaction time and accuracy, and, for Experiment 2, it provided a moderately good account of adaptation from one probability condition to another, but it could not correctly predict error reaction times nor fully account for the sequential effects of performance on one trial to the next.

Two GRAIN-Based Models

For one model, it was assumed that learning began at the beginning of the experimental trials and continued throughout the experiment. The model was presented stimulus–feedback sequences equivalent to those presented in the experiments, and it had to learn to perform the task from the feedback. For the other model, it was assumed that all learning had taken place before the experimental trials began; preexperiment training consisted of training the model for each possible stimulus to reproduce at output the probability that the stimulus was drawn from the high distribution. The first model could not correctly account for error reaction times, the sequential effects of one trial on the next, or the effects of switches in the stimulus probabilities (Experiment 2). The second model was generally successful with all response-time and accuracy measures and with sequential effects. The success of this model, like that of the diffusion model, depended on allowing variability in parameter values across trials. However, the model could not accommodate subjects' abilities to follow switches in stimulus probabilities.

In sum, the diffusion model succeeded extremely well, providing a coherent account of correct and error reaction times, reaction-time distributions, accuracy, sequential effects, and adaptation to switches in the probabilities of drawing stimuli from the high versus low distributions. It also provided an explanation of what drives the decision process. Although none of the connectionist models could give a satisfactory account of all the response-time and accuracy measures or of sequential effects and the effects of probability switching, our investigations lay out how they failed and provide a foundation for further theory development and evaluation.

Experiment 1

In Experiment 1, four subjects were tested in multiple sessions of the signal detection task in order to collect stable data for analyses of reaction-time distributions, error reaction times, and individual differences among the subjects.

Method

Subjects. The subjects were 4 Northwestern University undergraduates (3 men and 1 woman) who were paid $8 for their participation in each of 10 sessions. All had normal or corrected-to-normal vision.

Stimuli and apparatus. The asterisks were displayed in a 10 × 10 grid in the upper left corner of a VGA monitor, subtending a visual angle of 4.30° horizontally and 7.20° vertically. They appeared as light characters against a dark background, and were presented with high brightness and contrast and were clearly visible. The VGA monitors were driven by IBM AT-style microcomputers that controlled stimulus presentation time and recorded responses and response times.

The number of asterisks for presentation on a given trial was selected by randomly sampling from one of two discrete, approximately normal distributions with means 38 and 56 and standard deviation 14.4 (following Espinoza-Varas & Watson, 1994). The discriminability (d') between these distributions was therefore approximately 1.25. The two distributions crossed at the number 47; this number will be referred to as the "crossover" point for the two distributions. The display positions of the asterisks for a given trial were selected randomly from the possible 100 positions in the 10 × 10 character grid.

Procedure. Subjects were instructed that the number of asterisks on each trial was selected at random from one of two groups of numbers, a "low" group and a "high" group, and that the low group had fewer asterisks on average than the high group. The subjects' task was to decide whether the number of asterisks presented came from the low group, in which case they were to press the Z key on the computer keyboard, or the high group, in which case they were to press the *?* key. If a response was incorrect, the subject was informed immediately after the response. The subjects understood that they could not be completely accurate, that numbers from the middle of the range (e.g., 50) could have come from either distribution and that their task was to give their best judgment.

To provide some motivation to the subjects, a payoff scheme was used that awarded 4 points for every correct response and penalized 1 point for every incorrect response. Thus, for a block of 50 trials, a subject could earn as many as 200 points. Subjects were also encouraged to make their responses quickly, although they were told that their goal should be to maximize the total number of points earned over the course of the experiment. The points were not used to add to the payment rate for the experiment.

A trial began with the presentation of the asterisks. They remained on the screen until the subject responded, at which point the screen was erased. If the response was correct, a 700-ms waiting period ensued and then the asterisks for the next trial were presented. If the response was in error, the message "ERROR −1 POINT" appeared on the screen for 500 ms, followed by the next trial 700 ms later. Each block of 50 trials was completed in less than 5 min. Between each two blocks, the subject was encouraged to take a brief rest if he or she so desired.

Design. Each subject performed in 11 sessions (except Subject 1, who performed in 10 sessions) over approximately 3 weeks. Each session was composed of 24 blocks of 50 trials. Within a block, one half of the stimuli were sampled from the low distribution and one half were sampled from the high distribution. There were a total of 1,200 observations per session per subject. The first session was not used in any analysis (except for Subject 1), resulting in a total of 12,000 observations per subject. The first block of trials in each session was discarded from the analyses.

Results

In the data analyses, all of the trials with response times less than 200 ms or greater than 3000 ms were discarded (these constituted about .25% of the data).

The four subjects showed large individual differences in performance. One subject produced quite long reaction times (in the 400–800-ms range), another produced very short reaction times (in the 300–380-ms range), and the other two were intermediate. From a modeling perspective, this range of behaviors is a positive aspect of the data because it requires the models to have flexibility. If the models were too constrained, they might fit average data adequately but not the individual data of the more extreme subjects.

The presentation of the data is divided into three parts. First, it is shown that the probabilities of subjects' high and low responses followed, but were not the same as, the probabilities high and low stimuli. Also, three of the subjects showed sequential effects with the response on one trial being affected by the response on the previous trial. Second, responses generally slowed as the number of asterisks in the display was nearer the crossover point between the two distributions. However, across subjects, the relationship between correct and error reaction times varied. Third, the distributions of reaction times showed the typical skewed shape and their hazard functions rose and then either reached asymptote or fell slightly (as is typical of other tasks; see Luce, 1986).

Response probability and sequential effects. Figure 1 shows the probability of a low response for each subject as a function of the number of asterisks and the previous response. The probabilities fall smoothly from 0 to 96, and they cross the 50th percentile point close to the number 47, at which the low and high distributions crossed. Thus, the subjects performed without systematic biases.

The subjects differed in sequential effects. For Subjects 1 and 4, a response was a little more likely to be high if the prior response was high. In contrast, Subject 2 showed the reverse effect; when the prior response was high, there was a greater probability that the current response was low. None of the subjects showed any sequential effect that depended on the feedback given to the previous response, and Subject 3 showed no sequential effects at all. These individual differences (cf. Bertelson, 1961) present a challenge to models because the mechanism that produces sequential effects must be flexible enough to behave in opposite ways for different subjects.

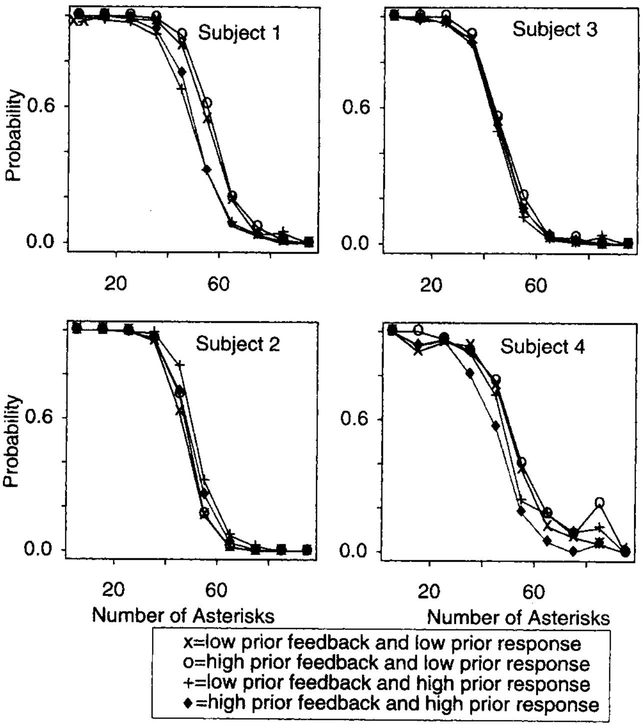


Figure 1. Probability of a low response for the four subjects in Experiment 1.

The fact that sequential effects were dependent on the prior response and not on prior feedback is consistent with most earlier findings with psychophysical tasks (Thomas, 1973, 1975; Treisman & Williams, 1984) and choice reaction time (Falmagne, Cohen, & Dwivedi, 1975; see Luce, 1986, chap. 7), although some studies, particularly in absolute identification (Ward & Lockhead, 1970), did find that feedback affected response probability. In the earliest investigations of signal detection paradigms, it appeared originally that any explanation of learning would have to take prior feedback into account (e.g., Kac, 1962), but Thomas (1973, 1975) showed that learning could be modeled by assuming criterion shifts toward the presented stimulus value so that learning did not depend directly on prior feedback. Thomas's account could also deal with paradigms in which feedback was not presented to the subject. Our experimental results are consistent with these early signal detection results and with the choice reaction-time results. Subjects knew that feedback was inconsistent and that for most stimuli the correct response was sometimes high and sometimes low. This, along with the large number of sessions tested per subject, probably explains why the feedback to the last response did not affect performance.

Response probability and mean reaction time. Because sequential effects in reaction time were small (on the order of 10–50 ms) relative to variability, reaction times were averaged over previous feedback and previous response. Figure 2 shows mean reaction time as a function of the displayed number of asterisks for

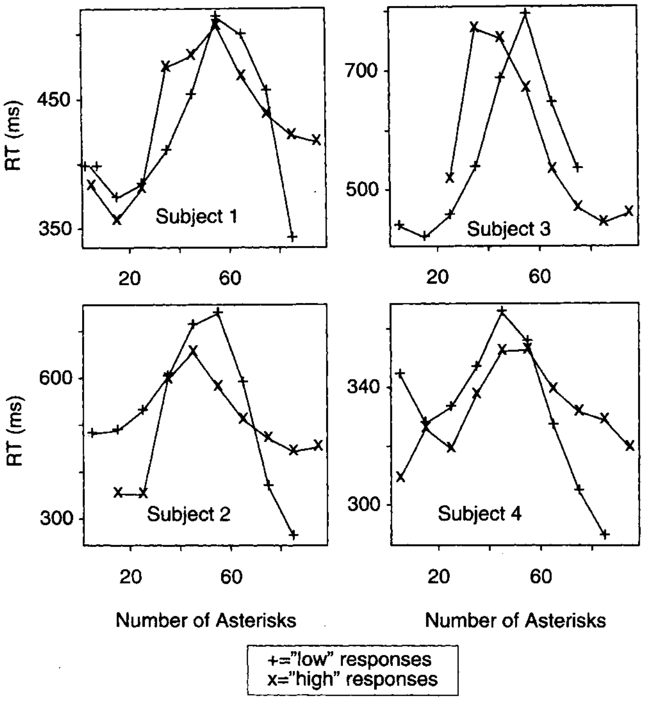


Figure 2. Mean reaction time (RT) for the four subjects in Experiment 1.

high and low responses. Generally, responses slowed as they neared the crossover point.

For purposes of exposition, we defined *error* responses according to the crossover point (47); low responses to numbers greater than 47 are labeled errors, and so are high responses to numbers less than 47. We used error as the label for these responses because it is a convenient way of describing them. A response of this type is not exactly an error, but neither is it the best response because it is less likely to be correct than the alternative. (Note that this definition does not correspond to the feedback that was given subjects; ERROR −1 POINT feedback was determined by the distribution from which a number was drawn, not by its position relative to the crossover point.) We use the error terminology for compactness of description throughout this article.

The subjects showed different patterns of error versus correct response times. For Subjects 1 and 2, errors for extreme stimulus numbers (e.g., numbers above 80 or below 20) were faster than correct responses for those numbers, whereas less extreme errors were slower than correct responses. But for Subject 4, errors were always faster than correct responses, and for Subject 3 errors were always slower than correct responses. This difference among subjects is the challenge to modeling outlined in the introduction; no model has yet been able to account for such variation while explaining the commonalities among subjects. In addition, no model has been able to account for a switch from slow errors to fast errors as response probability changes (Subjects 1 and 2).

A compact way to combine the reaction-time data and the response-probability data is to plot them jointly in a latency–probability function (Audley & Pike, 1965; Vickers et al., 1971). The reaction-time functions for high and low responses are reasonably symmetric about the crossover point (47), so they can be collapsed. So, for example, reaction times for low responses to 27 asterisks can be averaged with reaction times for the symmetrically equivalent high responses to 67 asterisks, and the probability of a low response to 27 asterisks can be averaged with the probability of a high response to 67 asterisks. Then the average reaction time can be plotted against the average response probability, as shown in Figure 3. Thus, the latency–probability function can be seen as a parametric plot where the parameter that varies along the plot is stimulus difficulty.

The different patterns of error versus correct response times show up in the degree to which the latency–probability functions are symmetric. Errors generally correspond to those responses with probability less than .5. A correct response with probability p corresponds to an error response with probability $1 - p$. For example, if the probability of a correct response is .8, the corresponding error probability is .2. If correct responses and their corresponding errors had the same response times, the latency–probability function would be a symmetric, inverted U-shaped function with a maximum at about .5. The function for Subject 3 is asymmetric, with errors always slower than their corresponding correct responses (see Figure 2). For Subjects 1 and 2, the functions are asymmetric, with errors slower than correct responses except that the most extreme errors are faster than correct responses. For Subject 4, the function is almost symmetrical, but errors are a little faster than correct responses.

Besides providing a summary of data, the shape of the latency–response probability function allows discrimination among various traditional sequential sampling models of reaction time (Audley & Pike, 1965; Vickers, 1979; Vickers et al., 1971). For example, a simple random walk model predicts a symmetrical inverted

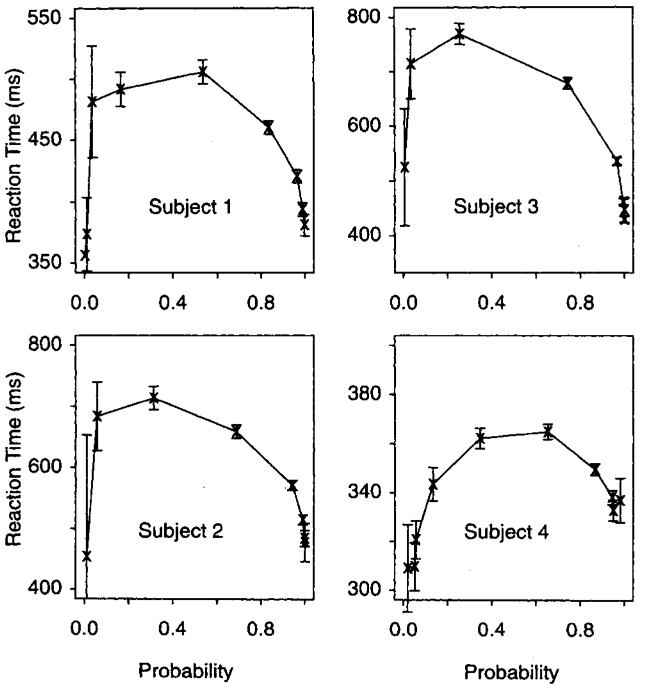


Figure 3. Latency–response probability functions for the data from Figures 1 and 2 for the four subjects in Experiment 1. The error bars represent 2 standard deviations.

U-shaped function, whereas recruitment models (LaBerge, 1962) with an absolute criterion predict an increasing function from fast correct responses to slow errors (Vickers et al., 1971). As Figure 3 shows, the data for none of the subjects conform to these predicted functions for the elementary versions of these models (Audley & Pike, 1965), although the pattern for Subject 3 can be handled by some models (Smith & Vickers, 1988; Vickers, 1979) and the pattern for Subject 4 by others (Laming, 1968; Link & Heath, 1975). These patterns of data are typical of those reported by Vickers (1979) and Vickers et al. (1971) in other paradigms.

Reaction-time distributions. Figure 4 shows representative reaction-time distributions for stimuli to which the low response was given over 90% of the time (e.g., 30–40 asterisks). As in almost all reaction-time research, the distributions are skewed to the right. Subject 2 showed a much wider central region with possible bimodality and a shorter tail than the distributions for the other subjects. Subject 2 also showed very large reductions in reaction time as a function of session, so data from this subject's first three and last three sessions were analyzed separately. Figure 5 shows this partition, with unimodal distributions that were much narrower and more skewed.

Figure 6 shows the reaction-time hazard functions derived from the reaction-time distribution histograms in Figure 4. The hazard function gives the likelihood for any point in time that the process will terminate in the next instant of time, given that it has not terminated before that time. Mathematically it is represented by $h(t) = f(t)/[1 - F(t)]$, where $f(t)$ is the density function at time t and $F(t)$ is the cumulative distribution function at t. When t becomes large, $F(t)$ approaches 1 so $[1 - F(t)]$ approaches zero, and the estimate of $h(t)$ becomes unstable (e.g., Bloxom, 1984, 1985; Luce, 1986) and severe oscillations in the estimate are common. The hazard function has been used as a method of testing detailed hypotheses about the distribution family from which a set of data arises (see Luce, 1986). Although hazard functions cannot

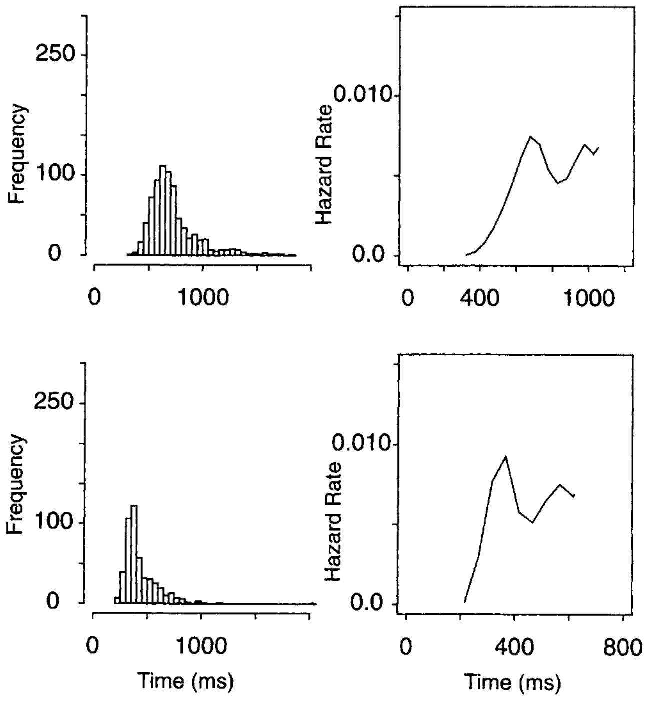


Figure 5. Reaction-time distributions and hazard functions for Subject 2 for the first three sessions (upper graphs) and the last three sessions (lower graphs) in Experiment 1.

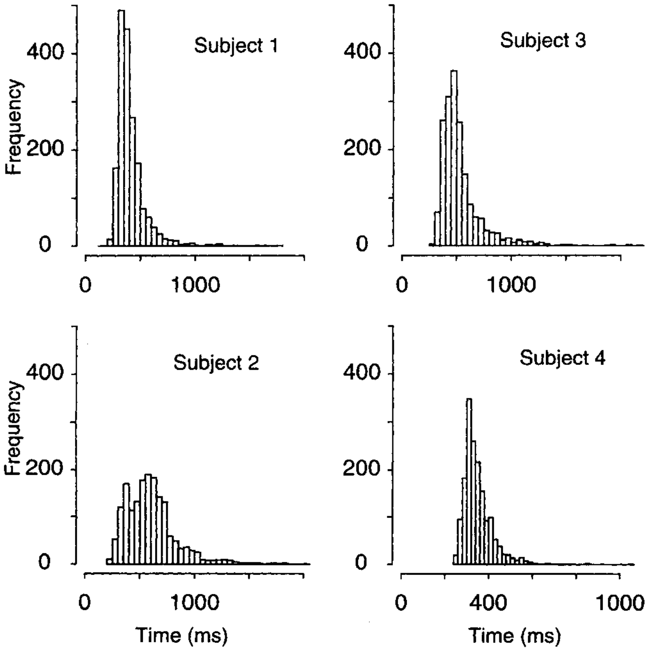


Figure 4. Typical reaction-time distributions for moderately high accuracy levels (for responses to stimuli in the 30–40 asterisk range) for the 4 subjects in Experiment 1.

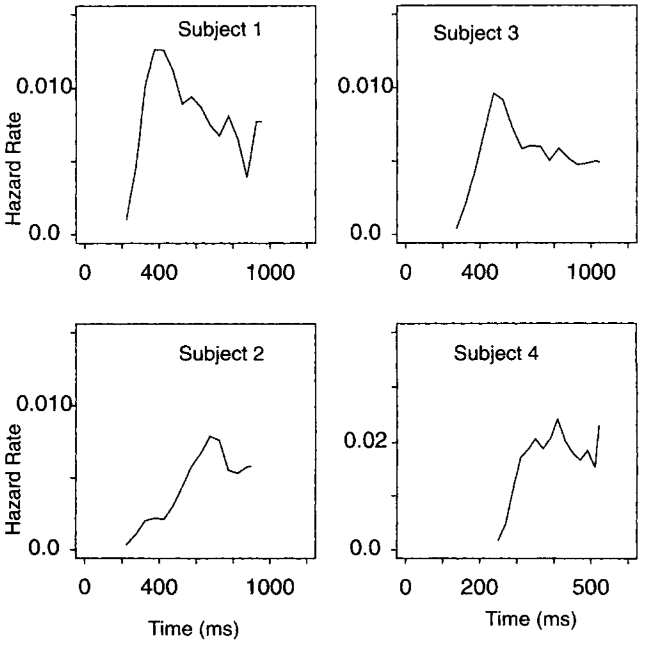


Figure 6. Hazard functions for the reaction-time distributions shown in Figure 4.

always be diagnostic of what distribution underlies an observed hazard function when there is variability across trials (see Burbeck & Luce, 1982; Van Zandt & Ratcliff, 1995), they are still useful in assessing model predictions. Here, hazard functions were calculated from the histograms (with equal width bins) in Figure 4. The hazard functions are reasonably stable up to 800 ms for Subjects 1, 2, and 3 and up to 450 ms for Subject 4 in Figure 6. In Figure 5, the hazard functions are stable up to 800 ms in the top panel and up to 600 ms in the bottom panel.

The hazard functions shown in Figure 6 are peaked, rising rapidly to a maximum and then falling gradually to the extreme tail where they become unstable. For Subject 2, Figure 5 shows the hazard functions for the first three and last three sessions of data, and these show similar trends to the other subjects.

Summary. The data from Experiment 1 present several targets for modeling: first, the shapes of the latency–probability functions and the different patterns of error versus correct response times shown by these shapes (Figure 3); second, the shapes of the reaction-time distributions for correct and error responses; third, the shapes of the hazard functions; and fourth, the sequential effects of the previous response on the current response (with no effect of feedback to the previous response). These aspects of the data are typically found in other cognitive paradigms. Therefore, success or failure of the models is significant not just for the signal detection paradigm but for application of the models to other cognitive paradigms.

There are intriguing variations among subjects shown in the differences in their mean response times and accuracy rates and in the differences in the speed of errors versus correct responses. Even though the task requires only that subjects learn to respond appropriately to stimuli drawn from simple probability distributions, the data challenge most current models because the qualitative behavior predicted by the models does not match the flexibility shown in the patterns of data across subjects. For example, some models predict fast errors, some predict slow errors, but none predict crossovers, and few models correctly predict the shape of reaction-time distributions. In the sections below, we show how the diffusion model takes up the challenge presented by the data, and then move to consideration of the connectionist models.

The Diffusion Model

The diffusion model was originally developed to explain the processes by which information is retrieved from memory over time (Ratcliff, 1978, 1980). It successfully fits data from binary choice recognition memory tasks (e.g., the Sternberg, continuous memory, prememorized list, and study–test paradigms). With no important modifications, it has also been applied to perceptual matching of letter strings (Ratcliff, 1981), to the varied and consistent mapping procedures with the Sternberg paradigm (Strayer & Kramer, 1994a, 1994b), and to new paradigms such as the speed–accuracy decomposition procedure (Meyer et al., 1988; Ratcliff, 1988). In all cases, it accounted for speed–accuracy relations, mean response times, and the shapes of reaction-time distributions. Models of this class have also recently received strong support from data from single cell recordings in monkeys (Hanes & Schall, 1996).

The one major failing of the diffusion model has been its inability to explain the relationship between correct and error response times. In Experiment 1, Subject 4 showed faster error responses than correct responses, Subject 3 showed slower error than correct responses, and Subjects 1 and 2 showed slower errors with intermediate stimuli and faster errors for extreme stimuli. Ratcliff (1978, 1981, 1988) stated that the diffusion model could produce only slower errors than correct responses and the predicted values were usually much slower than the data. We show here that the model can, in fact, accurately predict the varying patterns of error versus correct response times. The model failed to fit error reaction times in past applications because lack of computer power prevented an adequate search of the parameter space.

According to the diffusion model, information from a stimulus is continuously available and accumulated over time toward one of two decision boundaries. If the mean rate of accumulation is positive, the process generally moves toward the positive boundary, and if the mean rate is negative, it generally moves toward the negative boundary. There are two sources of variance in the diffusion model. First, the rate of accumulation, the drift rate, is noisy in that a particular path varies around the mean as it moves toward a boundary. Second, the mean rate of accumulation of information is different across the items in an experimental condition and across different instances of the same item. Ratcliff (1978) built this across-trial variance into the model to account for differences (variability) in memory strength (in application to recognition memory) for individual stimuli for a single experimental condition. However, the amount of across-trial variance was set at a constant in earlier fits to experimental data. As noted above, it was only when the parameter space could be automatically searched that we found that different values of this source of variance allowed the model to fit the different patterns of error reaction times. For more details on the qualitative behavior of the model, see the Appendix.

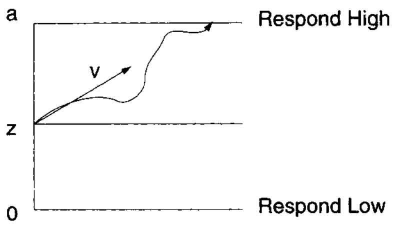
The diffusion model (and random walk and counter models generally) can be understood as an extension of signal detection theory to the time domain (e.g., Pike, 1973). Instead of sampling only once to determine a single value of strength for a stimulus, the diffusion model accumulates information continuously on the basis of repeated samples. The average across these repeated samples is the mean of the drift rate, and the standard deviation across trials in the average gives variability in the drift rate.

Fitting the Diffusion Model to the Data From Experiment 1

The parameters of the diffusion model are illustrated in Figure 7: The starting point and boundary parameters are z and a, the distances of the positive boundary and the starting point from the negative boundary (0), respectively. The amount of evidence needed to produce a positive response is $a - z$ and a negative response, $-z$. Because there were no apparent high or low biases in the subjects' response patterns and no reason to expect any from the experimental design, we set $z = a/2$ and eliminated z as a free parameter for the fits to the data from Experiment 1. The parameter v is the mean value of drift rate for the stimuli from an experimental condition. Each different possible number of asterisks in a stimulus display (1, 2, 3, . . . , 96) is a different experimental condition, so there are 96 different values of the parameter v. The two variability parameters are s, the standard deviation of drift within an individual process, and η, the standard deviation in mean drift rate across different trials of the same stimulus. The fourth parameter of the model is an encoding and response-time parameter T_{er} that represents the nondecision components of reaction time.

The parameter s (representing variability of drift within a trial) was set to the value 0.1 for all the fits of the diffusion model to the data; it was not a free parameter. This is because s is a scaling parameter; if s is altered, the other parameters of the model can be multiplied or divided by the ratio of the old and new values of s to produce exactly the same finishing time distributions and response probabilities as before s was altered. The value 0.1 was chosen because it is close to the value used in earlier applications of the model, and so comparisons can be made between parameter values here and in those applications (Ratcliff, 1978, 1981, 1988).

We did not fit the model to all 96 experimental conditions. Instead, we fit the model to three representative conditions, adjusting the parameters a, η, T_{er}, and three values of drift rate v (one for each condition) to produce the best fit of model to data from the three conditions. Then, with a, η, and T_{er} held constant, v was varied to produce predictions for all the other 93 conditions (the data were actually collapsed into 10 groups). The three conditions used in fitting were chosen to represent widely spaced parts of the latency–response probability function (Figure 3). For example, for Subject 1, the values of response probability chosen were 0.965, 0.463, and 0.143. Each of these three values actually corresponds to two sets of reaction-time data, the number for which the probability of a high response equaled the chosen value and the number for which the probability of a low response equaled the chosen value. The fitting program adjusted the three values of v plus the three other parameters (a, η, and T_{er}) to minimize a sum of squares using a standard function minimization routine. The data for the different subjects were fit individually, so the three values of v plus the other three parameters all were free to vary across subjects. The parameter estimates are shown in Table 1.



Parameters of the Diffusion Model:
a = Boundary position
z = starting point = a/2
v = mean drift rate, one for each condition
s = standard deviation in drift within a trial
T_{er} = encoding and response time
η = standard deviation in mean drift rate from trial to trial (drift is N(v,η))
s_z=standard deviation in starting point (starting point is N(z,s_z))

Figure 7. The diffusion model and parameters of the model.

Table 1

Parameters of the Diffusion Model Used in Fitting for the Four Subjects in Experiment 1 and 2 Subjects in Experiment 2

Subject (S)	a	z	T_{er}	η
Experiment 1: S1	.115	a/2	.256	.112
Experiment 1: S2 (first 3 sessions)	.151	a/2	.323	.088
Experiment 1: S2 (last 3 sessions)	.141	a/2	.335	.174
Experiment 1: S3	.150	a/2	.313	.142
Experiment 1: S4	.065	a/2	.266	.055
Experiment 2: S1	.117	.040	.206	.082
Experiment 2: S2	.103	.039	.306	.081

Note. a, z, T_{er}, and η are parameters.

The sums-of-squares function for minimization was constructed as follows.[1] First, the empirical reaction-time distributions were fit with an ex-Gaussian distribution, that is, a convolution of normal and exponential distributions. The ex-Gaussian has been shown to provide a good summary of empirical reaction-time distributions (Ratcliff, 1978, 1979; Ratcliff & Murdock, 1976), and its parameters have been used to describe the shape of the distribution. Theoretical distributions were then generated by the diffusion model, and these theoretical distributions were also fit with an ex-Gaussian distribution. The two parameters (μ and τ) of the ex-Gaussians served as a meeting point between the empirical data and the theoretical predictions from the model (μ roughly represents the position of the leading edge of the distribution, and τ represents the extent of the tail of the distribution). The sum-of-squares function was the sum of squared differences between the theoretically derived and empirically derived values of the ex-Gaussian summary parameters plus the sum of squared differences in the theoretical and empirical values of response probability (all weighted by standard errors). The fitting routine minimized the sums of squares as a function of the diffusion model parameters (see the Appendix for a full presentation). (The ex-Gaussian has a third parameter, σ, which roughly specifies the rise in the leading edge of the distribution, but it is not needed because the diffusion model produces a rise in the reaction-time distribution that is close to the rise observed in the experimental data.) All fits of the model shown in the figures are direct fits of the model to the data. In more recent work, we have moved to fitting the reaction-time distributions directly using quantiles of the distributions. The obtained fits are not different in the two procedures.

As pointed out, the three values of v for each subject were merely representative of all of the 96 experimental conditions and served the purpose of summarizing the range of data and allowing the other three parameter values, a, η, and T_{er} to be fixed for the subject. To sweep out all the conditions, v must be varied from some very low value to some very high value. It turned out that all of the conditions were accommodated by v ranging from $-.4$ to $+.4$, where a drift rate of $-.4$ corresponded to less than 20

[1] Note that setting up a successful run of the fitting process usually requires one or more runs much of the way through the process before a result can be obtained because the program is quite sensitive to the starting values of the parameters. The initial parameter values have to be close to the final values or else estimates start to diverge, numerical overflow or underflow occurs, and the program terminates.

asterisks and a probability of greater than .98 that 20 asterisks was drawn from the low distribution, and a drift rate of 0 corresponded to 47 asterisks and a probability of .5 that 47 asterisks was drawn from the low distribution. With v varying across this range and the other three parameters at their fixed values, the diffusion model had to predict, for all 96 conditions, all the standard speed–accuracy measures: the probability of a high response, the speed of high and low responses, the shapes of distributions of response times, and the hazard functions of the distributions. In the next sections, we show the model's success with these predictions.

Latency–Response Probability Functions and Error Reaction Times

The first test of the diffusion model was to examine whether it accurately fit the latency–response probability functions displayed in Figure 3. These functions are shown again in Figure 8 along with the fits of the diffusion model. The model fits the data well with only the single parameter v varying (from $-.4$ to $+.4$). The model explains the large differences in average reaction time between Subjects 3 and 4 as the result of differences in the a parameter, that is, differences in boundary positions (see Table 1). The error bars shown in Figure 8 represent plus or minus 2 standard deviations in reaction time. The fits are good, with only about 3 or 4 of the 40 data points lying more than 2 standard deviations outside the theoretical functions. Although different values of drift rate are needed to fit each individual condition in the experiment, the shape and location of the latency probability function shown in Figure 8 is a function of only the three parameters a, η, and T_{er}.

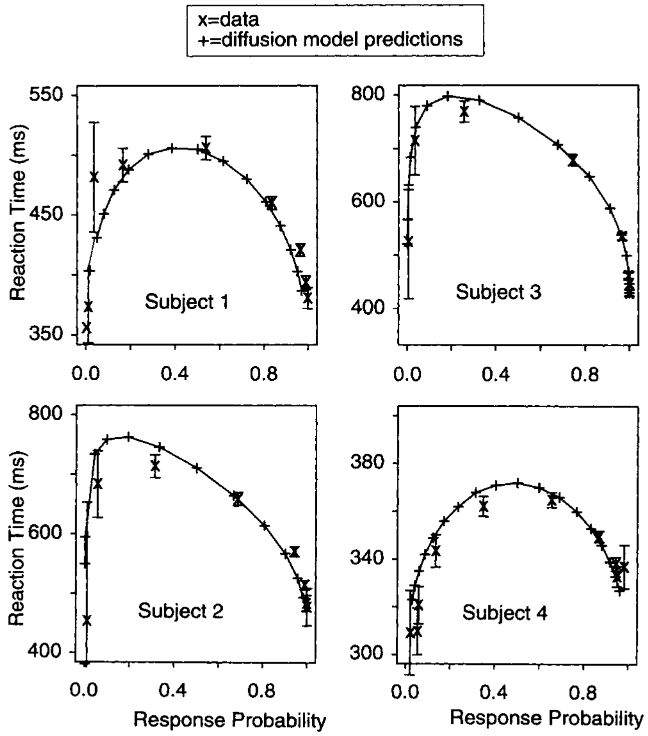


Figure 8. Diffusion model fits to the latency–response probability functions for Experiment 1. The error bars represent 2 standard deviations.

The different shapes of the latency–response probability functions for the different subjects reflect different patterns of error versus correct response times. The model shows flexibility in accounting for these different shapes with different values of the three parameters a, η, and T_{er}. For example, the model was able to fit the near symmetry of the function for Subject 4 as well as the extreme asymmetry of the function for Subjects 2 and 3. The main determinant of symmetry–asymmetry was the size of η, the standard deviation in mean drift values from trial to trial (see Table 1 for values).

To see how the standard deviation in mean drift across trials determines the shapes of latency–probability functions, first consider a diffusion process with the starting point halfway between the boundaries and with no variability in drift across trials; then error reaction times are the same as correct reaction times. When variability in drift rate across trials is introduced, then responses are a weighted average of reaction times for the different drift rates (weighted by the probability of each response). Table 2 illustrates this weighting. Columns 2 to 4 show correct and error reaction times and response probability. To illustrate variability in drift, we average each row with adjacent rows (so the average reaction time or accuracy for drift variability v is the average of rows with constant drift $v - .05$, v, and $v + .05$ with fixed v). The results of this averaging are shown in columns 5 to 7. For example, averaging the three constant drift values $v = -.10$, $-.15$, and $-.20$, the average error response time 663 ms was computed from the error response times for the three v values (694 ms, 599 ms, and 535 ms), each weighted by their probability (.076, .023, and .007, respectively): $(535 \times .007 + 599 \times .023 + 694 \times .076)/(.007 + .023 + .076) = 663$ ms. The corresponding average correct response time (weighted by probabilities .993, .977, and .924, respectively) was 607 ms. Graphing the probabilities of the responses (the probability correct in column 7 and 1 minus the probability correct for errors) against the averaged correct and error response times (columns 5 and 6) produces an asymmetrically shaped latency probability function such as those from Subjects 1, 2, and 3.

The diffusion model fails to accurately capture the shape of the latency–response probability function only for Subject 4, for whom the model shows error responses never faster than correct responses, in contrast to the data in which error responses were systematically faster than correct responses. There is an explanation of fast errors that is often given for random walk models, and that is that the fast errors result from variability in the starting point of the walk (e.g., Laming, 1968, see also Ratcliff, 1981). When the starting point is near a boundary, then the probability of reaching that boundary in error (with a fast reaction time) is greater than when the starting point is further away from that boundary (slow errors with lower probability). Averaging faster errors (weighted with higher probability) with slower errors (weighted with lower probability) produces faster errors on average than the case where the starting point, z, is constant.

To test whether this explanation was tenable for Subject 4, we compared the model's prediction using a single starting point ($z = a/2 = 0.026$; see Table 1) to predictions using a normal distribution of starting points with standard deviation $0.2z$ (.0052). The result is shown in Figure 9. With a standard deviation of 20% in

Table 2
Correct and Error Reaction Times for the Diffusion Model for Fixed Drift Rate and a Distribution of Drift Rates (Using Parameter Values From Subject 3 Fits)

	Reaction time (RT) and probability (Pr) are for fixed drift rate ($\eta = 0$)			RT is weighted average of RT in column 2 for $v - .05$, v, $v + .05$. Corrects are weighted by Pr correct and errors by $1 -$ Pr correct		
v	RT correct (ms)	RT error (ms)	Pr correct	RT correct (ms)	RT error (ms)	Pr correct
.05	812	812	.223	—	—	
0	875	875	.500	833	833	.500
−.05	812	812	.777	777	840	.734
−.10	694	694	.924	694	769	.893
−.15	599	599	.977	607	663	.965
−.20	535	535	.993	542	578	.989
−.25	492	492	.998	496	519	.997
−.30	462	462	.999	—	—	

Note. A latency probability function would be constructed by plotting reaction times in columns 5 and 6 against Pr correct in column 7 for correct responses and (1 − [Pr correct]) for error responses. v = drift rate.

the value of z, error reaction times are speeded up exactly enough to match the data. For the other subjects, z is larger so that variability in the range ± .0052 makes errors to extreme stimuli faster (and so produces fast errors for extreme values of accuracy for Subjects 1 and 2, which is in better accord with the data) but has little effect on response times for intermediate errors (which are slow errors).

In earlier research, Ratcliff (1978, 1981, 1988) had stated that the diffusion model could not correctly predict error reaction times. For recognition memory paradigms, error reaction times are generally slower than correct reaction times, but not as slow as the diffusion model seemed to predict. However, Figure 8 shows that, for the signal detection paradigm, reaction times are reasonably well fit over the whole range of response probability values including values below .5, that is, including values for which the responses were errors (except for the small deviations for fast error reaction times noted above for Subject 4). The reason that the diffusion model now predicts error reaction times correctly is that the minimization program was able to adjust the drift variability parameter (η). For practical reasons (fitting by hand because of a lack of computational power), Ratcliff (1978) had fixed this parameter, adjusting only the other parameters of the model. Refitting some of the old recognition memory data with the new minimization program showed that good fits for error reaction times can be obtained for the recognition memory paradigms used in Ratcliff (1978), as well as for the data reported here.

It is important to note that it is variability in drift and variability in the starting point of the diffusion process that allow the model to predict the complicated pattern of error and correct reaction times. Van Zandt and Ratcliff (1995) showed that seemingly decisive tests between models could become indecisive once variability in the parameters of the models across trials was introduced. Our results go one step further by showing that patterns of data that could not be explained by the diffusion model using a fixed mean drift rate (i.e., $\eta = 0$) and a fixed starting point can be explained with variability in those parameters across trials.

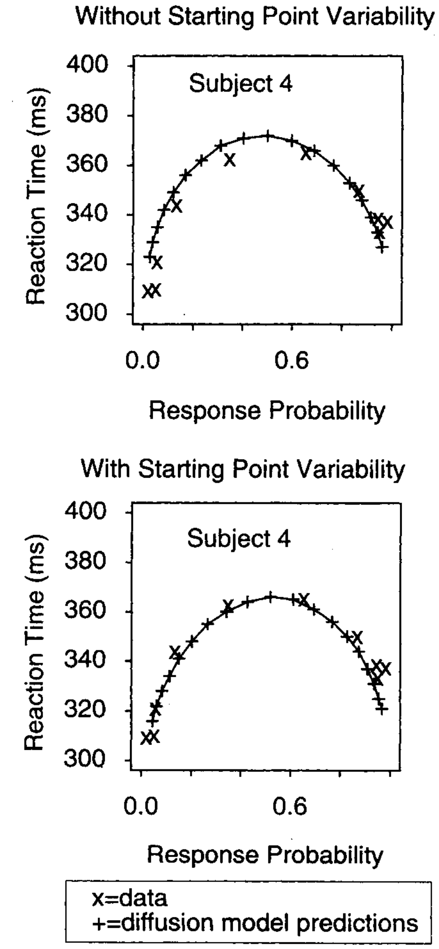


Figure 9. Latency–probability function predictions for the diffusion model without starting point variability (top) and with starting point variability (bottom) with standard deviation = .2z for Subject 4 in Experiment 1.

Reaction-Time Distributions and Hazard Functions

The second test of the diffusion model was to examine whether it could accurately predict the shapes of reaction-time distributions. Figures 10 and 11 demonstrate that it can. The figures show the empirical distributions and model predictions for two conditions per subject, one condition for which response probability was high (e.g., near 0.95) and another condition for which probability was lower, in the 0.5–0.7 range (the model makes the same predictions for high as low responses at any probability level because the response boundaries are symmetrical in these fits). The hazard functions (Figures 12 and 13) showed functions either rising to asymptote or rising to asymptote then falling slightly. The functions are shown for only two conditions because the distribution shapes and hazard function shapes did not deviate from the illustrated patterns across conditions.

The accuracy of the diffusion model's predictions is especially noteworthy because fitting the diffusion model used only the two parameters of the ex-Gaussian to summarize distribution shape. Also, there were no additional free parameters used in fitting the distributions beyond a, η, T_{er}, and the values of mean drift rate v.

Chi-square goodness-of-fit values for the eight distributions are shown in Figures 10 and 11. They were computed using observed and expected numbers of counts in the histograms in Figures 10 and 11. For three of the distributions, the theoretical and empirical distributions are significantly different from each other, but for five they are not. This is comparable with the goodness of fit of other models to reaction-time distributions (e.g., Ratcliff & Murdock, 1976). There are several possible reasons for the less than perfect fits, including variations in the behaviors of subjects across sessions (e.g., a slow day or a fast day), long-term practice effects, and variability in the nondecision component of processing (T_{er}). These factors would all produce a less abrupt than predicted rise in the leading edge of the distribution, which is where most of the misses occur. But overall, these fits are very good because there is no other systematic misprediction.

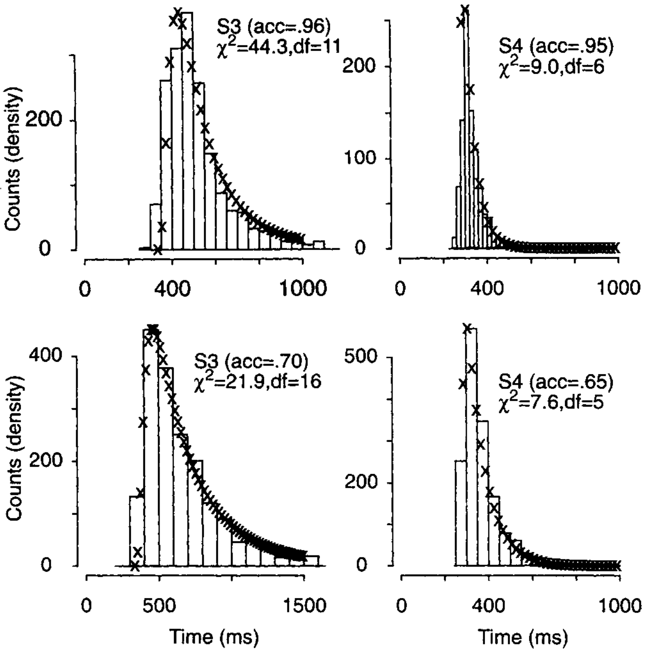


Figure 11. Diffusion model fits to the reaction-time distributions for Subjects 3 and 4 (S3 and S4) in Experiment 1. The Xs are the model predictions. acc = accuracy.

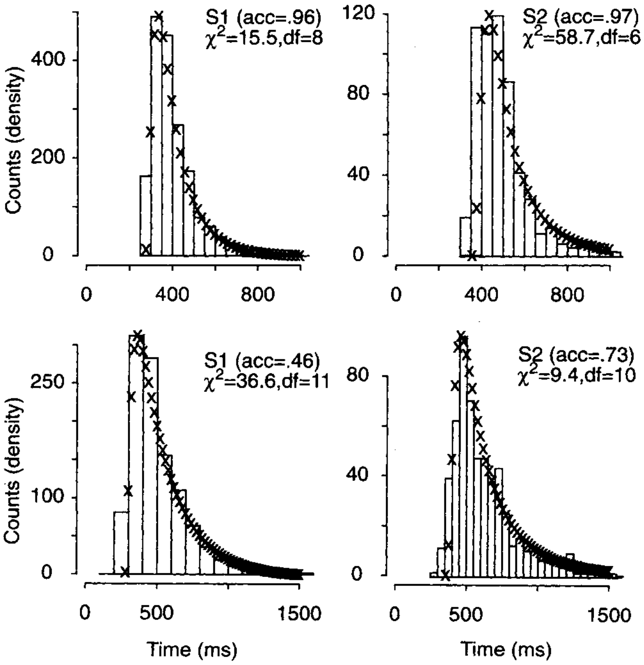


Figure 10. Diffusion model fits to the reaction-time distributions for Subjects 1 and 2 (S1 and S2) in Experiment 1. The Xs are the model predictions. acc = accuracy.

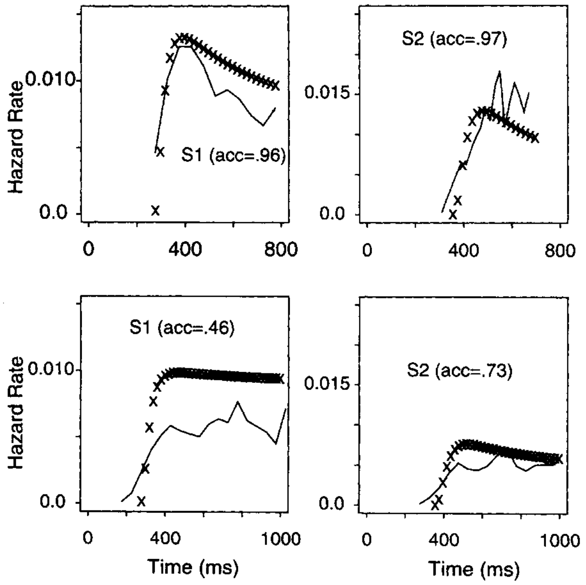


Figure 12. Diffusion model fits to the reaction-time hazard functions for Subjects 1 and 2 (S1 and S2) in Experiment 1. The Xs are the model predictions. acc = accuracy.

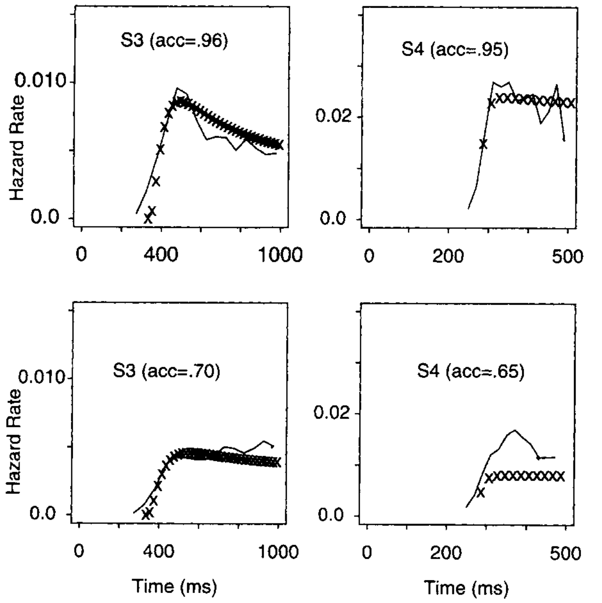


Figure 13. Diffusion model fits to the reaction-time hazard functions for Subjects 3 and 4 (S3 and S4) in Experiment 1. The Xs are the model predictions. acc = accuracy.

Stimulus Probability and Drift Rate

For recognition memory, drift rate in the diffusion model corresponds to the strength or familiarity of an item in memory (Ratcliff, 1978). An item that is more familiar might be one that is more meaningful to the subject or one that is studied longer. But what dimension does drift rate correspond to in the signal detection paradigm?

What subjects ought to do is make their choice based on the information that they have from the experiment. The one piece of information that they have (after some number of trials, each with feedback) is the probability with which a number of asterisks was drawn from the high distribution versus the low distribution. One possibility then is that the subjects base their choices on this probability. From a long tradition of probability learning literature (see Atkinson et al., 1966; Estes, 1957, 1964; and, especially, Estes, 1995, for a detailed review), we know that subjects can do this, at least in many situations. But there is nothing in the data from Experiment 1 that directly suggests that the subjects in the experiment were probability matching. For instance, the response probabilities shown in Figure 1 do not directly correspond to the probability that some number of asterisks was high or low: The response probability functions are too abrupt when compared with the stimulus probability function. Instead, in the diffusion model, it might be that the mean drift rate is derived from stimulus probability. In other words, the mean drift rate for a given number of asterisks might reflect the probability that the number was drawn from one or the other of the two possible distributions, the high distribution or the low distribution.

We discovered this possibility, not by intuition, but when we plotted the drift rates for the four subjects as a function of experimental condition (i.e., number of asterisks) and compared the drift rates with the probabilities with which each possible number of asterisks was drawn from the high versus low distribution in a typical experimental sequence. The subjects' functions are shown in Figure 14. The probability values for a typical experimental sequence were generated from the same algorithm that was used to generate stimuli in the experiment, with the length of the sequence the same as the total number of trials a subject would receive in all of their sessions combined. For each possible number of asterisks, N, the probability that N asterisks was drawn from the high distribution was plotted in Figure 14. In order to plot the probabilities on the same scale as the subjects' drift rates, the probabilities were transformed to lie between −.4 and +.4 by subtracting 0.5 and multiplying the result by 0.8, so that a probability of .5 (middle of the range) corresponded to a drift rate of 0 (middle of the range). There are two dramatic results: First, the four individual subjects have very similar drift rates across conditions despite their large differences in other performance measures. Second, the subjects' drift rates correspond very closely to stimulus probability.

These were unexpected results for two reasons. First, the four subjects performed differently from each other in radical ways, in terms of overall reaction times, speed–accuracy tradeoffs, and in the parameters of the diffusion model other than drift rate. Examination of their individual reaction-time data gives no clue to any significant underlying similarities or invariances across subjects. Second, from the plot of drift rates in Figure 14 and the plot of response probabilities in Figure 1, drift rates do not clearly map into response probability: The drift rate functions are gradual over the whole range of numbers of asterisks, whereas the response probability functions have a steeper climb in the middle of the range and asymptote much more quickly. Yet, according to the diffusion model, the subjects were all basing their decisions on the same underlying variable. What Figure 14 shows is that stimulus probability, mapped through the diffusion process, produces the observed response-probability and reaction-time data. Similar

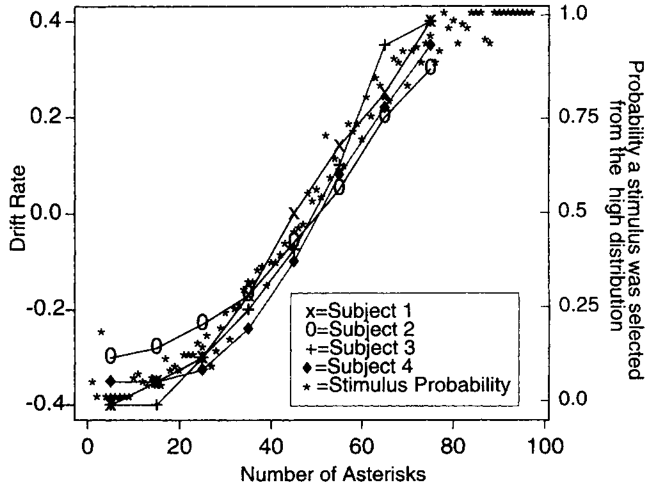


Figure 14. Diffusion model drift rates for the four subjects. The asterisks represent the probability that the number of asterisks came from the high stimulus distribution, transformed to the range −.4 to +.4 from the probability range 0 to 1 (subtracting .5 and multiplying by .8). The probability function (asterisks) has variability derived from running simulations of about as many trials as for one subject for all sessions.

drift-rate–stimulus-probability correspondences have been found by Ratcliff and Rouder (1998) across a range of experimental procedures including brightness discriminations, red–green color discriminations, and same–different brightness judgments.

Given the relationship between drift rate and stimulus probability, the number of parameters for the diffusion model can be reduced. The drift rates themselves, the 96 values of v, are no longer needed as parameters of the model. They can be replaced by the transformed probability values. So the three v parameters that were used for each subject in the original fit of the diffusion model can be replaced by the two parameters needed to transform probability to drift rate. This means that the complex pattern of results including response probabilities, the shape of the reaction-time distributions, and the hazard functions for correct and error responses can be modeled accurately by the diffusion model with only six free parameters per subject: two parameters that scale probability to drift and a, η, variability in starting point s_z, and T_{er}.

Although these results at first appeared to us to implicate stimulus probability as the function driving performance, we cannot rule out another function: distance from a criterion. Subjects could be setting a criterion on the dimension of numerosity and responding high if the number of asterisks in the stimulus was greater than the criterion and responding low if the number of asterisks in the stimulus was less than the criterion. Variability in encoding of the stimuli (and possibly also variability in the criterion setting) would lead to a gradual transition in mean drift rate from high to low. In the domain of categorization research, exemplar-based models and distance from criterion models have been compared (e.g., Maddox & Ashby, 1993; Nosofsky & Palmeri, 1997), and it has been concluded that there is a great deal of mimicking between the two classes of models. Early exemplar-based models essentially predicted probability matching as the function guiding responses, but individual subjects produced more deterministic responding (sharper functions, e.g., Figure 1) than were predicted. More recent exemplar-based models have used a random walk as the decision process and shown how gradual stimulus-probability functions could be mapped into sharper response-probability functions by summation in the random walk (a well-known property of many sequential sampling models). This situation in the categorization literature, mimicking between the two classes of models, applies in our research as well: The function driving the diffusion process could be either stimulus probability or distance from a criterion. Resolution of this situation will require experiments that produce differential predictions for the two kinds of models.

One of the reviewers asked why it was not expected a priori that stimulus probability would be the function driving drift rate in the diffusion model; after all, the random walk process is a probabilistic one. There are several reasons why this relationship would not necessarily be expected. First, only in the random walk (the discrete analog of the diffusion process; Feller, 1968; Ratcliff, 1978) can the probability of a step toward one response boundary or the other be directly derived from stimulus probability. Taking the limit in the random walk to produce the diffusion process, the step probability has to approach .5, not stimulus probability. Second, even in the discrete random walk, the probability of taking a step toward one boundary versus the other cannot be stimulus probability in many experimental paradigms. For example, in a two-choice reaction time task, feedback would always correspond to the stimulus presented. Unlike our asterisks task, feedback is always accurate, so that stimulus probability is 1 for each stimulus. If this stimulus probability was the probability of a step toward the correct boundary, then that probability would be 1, and each step in the process would be deterministic, so all responses would terminate in the same number of steps and all responses would be correct. Third, in other random walk models, the rate of approach to the boundaries is not derived from stimulus probability. For example, Link and Heath (1975), in their random walk model, assumed that the distance moved toward one or the other boundary in one step was a function of distance between the transduced stimulus and a noisy criterion. Fourth, we performed a pilot experiment with two-digit numbers as stimuli instead of arrays of asterisks. We used the same feedback and stimulus selection scheme as in Experiment 1. Results showed that subjects set a fixed criterion around the number 55; stimuli higher than 59 and lower than 50 all produced approximately the same mean reaction times and response probabilities, near ceiling. In fits of the diffusion model to this data, drift rates would not be a function of stimulus probability.

In sum, although a reasonable hypothesis to entertain is that drift rate is a transformation of stimulus probability, this does not follow directly from the theory of the diffusion model, other random walk models make other assumptions, and other data sets contradict the hypothesis (but see Ratcliff & Rouder, 1998).

Sequential Effects

The one aspect of the data left to discuss are the sequential effects from one response to the next. It turns out that the diffusion model can account for these effects in the same way that it accounts for how subjects accommodate to changes in the probabilities with which stimuli are given high versus low feedback. Experiment 2 examines in detail the consequences of such switching, and so we postpone discussion of sequential effects until after presentation of the results of that experiment.

Summary

The success of the diffusion model in this application is that it gives a good account of the data parametrically and, at the same time, offers insight into how the stimuli in the signal detection paradigm controlled responses, what aspects of performance were common across subjects, and what aspects were different. There was no way to see directly from the data that the subjects took stimulus probability into account in their performance, and no way to see exactly what the relationship was between stimulus probability and response probability or response time. The model shows how stimulus probability (or distance from a criterion) can be transformed into drift rate that in turn, acting through the mechanics of the diffusion model, can produce the whole range of characteristics of the data. At the same time, the model shows how the different individual subjects can show quite different response-time–accuracy profiles yet still all be governed by the same variable, stimulus probability.

The fact that the diffusion model fits the data well raises the issue of falsifiability. If only mean reaction time and response probability were taken into account, it might be difficult to falsify the model. But once the shapes of reaction-time distributions are taken into account, falsification would be very easy: The model has to predict that reaction-time distributions are skewed to the right, not skewed left or bimodal, and it places tight limits on how much minimum reaction times can differ between error and correct responses for a particular experimental condition. The model also makes a strong prediction for experimental situations in which the boundary positions are fixed, as they must be, for example, in an experiment in which the manipulation of variables within a trial makes it impossible for subjects to anticipate which condition is being tested and thus impossible for them to shift boundaries between conditions (see Ratcliff, 1978, Experiment 1). In this situation, the model predicts that as drift rate varies across experimental conditions, increases in reaction time come from spread in the tail of the reaction-time distribution with relatively little change in the fastest responses (in ex-Gaussian terms, this would correspond roughly to τ increasing 3 or 4 times more than μ).

Another way to look at the falsifiability issue is to consider how a failure of the diffusion model to fit data might be interpreted. The model predicts a large spread in the tail of the distribution when shifts in the leading edge are small. If shifts in the leading edge are too large relative to spread in the tail, this might signal that another stage of processing had been inserted in one condition of an experiment relative to another. Such an inserted stage should also show up as a shift in the onset of growth of accuracy in response signal data (e.g., see McElree & Dosher, 1993, for discussion, and Hacker, 1980; Hockley, 1984; Muter, 1979, for examples of shifts in the leading edge of reaction-time distributions). Thus, although the diffusion model on initial inspection might seem to be hard to falsify, it does have tight constraints, especially for predictions about the shapes of reaction-time distributions.

The success of the application of the diffusion model to the signal detection paradigm is also a significant step forward for traditional modeling. For the first time, the description offered by a model of how information is accumulated over time leads to an accurate and unified account of both error and correct response time, including the shapes of the reaction-time distributions and their hazard functions. In the past, even the most successful models could not simultaneously and accurately account for all of these aspects of the data.

Alternative Standard Reaction-Time Models

The diffusion model is one of the class of sequential sampling models that includes random walk models, counter models, and runs models (see Luce, 1986). Of these models, the random walk models (discrete versions of the diffusion model) are of the same family as the diffusion model and show much the same behavior as the diffusion model. The standard random walk model predicts that the mean number of steps to cross a boundary is the same for correct and error responses (if the starting point is equidistant from both boundaries). In order to account for choice reaction-time data, Laming (1968) added starting point variability to the random walk to produce error reaction times faster than correct reaction times. Link (1975) and Link and Heath (1975) allowed the step size in the process to be sampled from a nonnormal distribution, which, depending on the distribution, allows error responses to be either faster or slower than correct responses. Neither of these models could produce the crossover between correct and error reaction times obtained for some subjects in Experiments 1 and 2. However, these models might be adapted using variability in parameters across trials to produce the same success as the diffusion model.

The recruitment model of LaBerge (1962) assumes that information (e.g., features) from a stimulus is accumulated to a fixed criterion (in two-choice tasks, to one of two criteria). This model differs from the random walk model in the use of absolute rather than relative criteria. The fixed-criterion recruitment model makes several predictions that are at odds with experimental data: It predicts that reaction-time distributions become less skewed as the criterion number of counts is increased. Also, it predicts that there is a maximum reaction time (number of counts) for any process, and this maximum is one less than the criterial number of counts for each counter plus one. The experimental data suggest that there is no fixed upper limit on reaction time. The recruitment model also usually predicts negatively skewed error reaction-time distributions, contrary to experimental data.

Accumulator models (Smith & Vickers, 1988; Vickers, 1970, 1979) are variants of the counter models that assume evidence to be continuous (rather than discrete counts) and time steps to be discrete. Smith and Vickers extended the early accumulator model by assuming that time steps have an exponential distribution. This modification produces correct and error reaction-time distributions that are positively skewed. However, if the criteria are increased, the reaction-time distributions become more normal, contrary to experimental data. Smith and Vickers assumed that the response criteria are variable (like the starting point of the random walk in Laming's model and the diffusion model) and if variability increased with criterion position (or the distribution of criterion positions was skewed), then the predicted reaction-time distributions would become more skewed and the model might be capable of fitting the experimental data reported here. But it is unclear whether the parameter invariance (drift rates, parameters other than boundaries in speed–accuracy manipulations; Ratcliff & Rouder, 1998) found for the diffusion model could be obtained by the accumulator model.

Another popular alternative to sequential sampling models is the class of strength–latency models, or distance-from-a-criterion models. These models assume that reaction time is a function of distance of the stimulus value from a decision criterion. The most popular instantiation of this class is a model in which reaction time is an exponential function of distance from the criterion (Ashby & Maddox, 1994; Murdock, 1985; see also Vickers, 1979, p. 144). Although this model can make predictions for accuracy and correct reaction times that are consistent with data, the model makes the prediction that error reaction-time distributions are negatively skewed (see Figure 15). Also, the model predicts that errors can only be slower than correct responses because the peak of the strength distribution that produces most of the correct responses is further away from the criterion than the tail of the distribution that produces errors.

Connectionist Models

One of the appealing features of connectionist models is that they provide explicit mechanisms for learning. Given many instances of stimuli, each with feedback, the models can learn mappings from stimuli to responses; these mappings lead to predictions about the probabilities with which different responses will be given to different stimuli. Because, in many connectionist

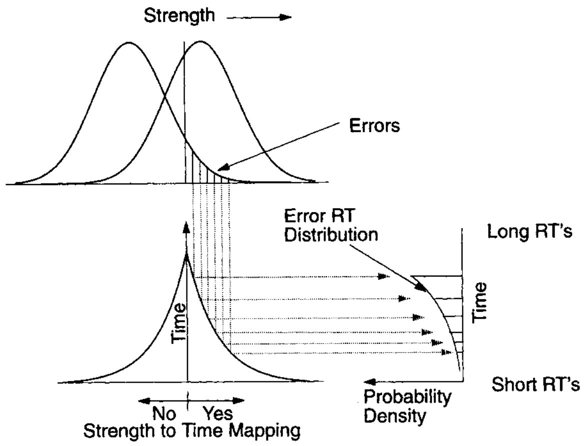


Figure 15. An example of the strength–latency model mapping from strength through an exponential latency function to the reaction-time (RT) distribution.

models, the process by which the response to a stimulus is chosen involves iterative recycling of activation, it seems natural that the models should also generate predictions about response-time measures. However, the predictions for response time generated by earlier models were often wrong, and the predictions of more recent models have been largely untested. We begin discussion of the treatment of response-time measures in the connectionist domain with a review of earlier models and their problems, and then proceed to test two models based on McClelland's GRAIN framework (McClelland, 1993; Movellan & McClelland, 1991, 1993) and also Anderson's (1991) BSB model, using the response time and response probability data from Experiment 1.

The Cascade Model (McClelland, 1979)

The cascade model was one of the first models to examine the possibility of information flowing continuously through discrete stages of processing. The key assumption was that one stage of processing could begin before the previous stage had terminated. The model's ability to account for mean response-time data with this cascade notion posed a challenge to additive factors logic (Sternberg, 1969), which had been used to separate the effects of different variables on different stages of processing. However, the cascade model was eventually shown to have shortcomings; most seriously, it predicted that on some proportion of trials, processing would never terminate (Ashby, 1982). Even with added assumptions by which processing could be terminated, the model could not adequately predict reaction-time means and variances (Ashby, 1982). However, the model did set the stage for other dynamic models based on activation flow through a series of stages.

The Interactive Activation Model (McClelland & Rumelhart, 1981)

The interactive activation model was designed to explain performance in simple letter and word identification tasks. It was an immediate successor of the cascade model, but it was not designed to predict response times for decisions, only the probabilities of making different decisions. In this model, information about words is represented in a network of three levels of nodes: letter feature (line segment) nodes, letter nodes, and word nodes. The features of a stimulus word are input at the feature level and then activation flows among the nodes of the different levels until the amount of activation asymptotes or, in paradigms in which the stimulus is masked, it reaches a maximum. The asymptotic or maximum amount of activation at the word or letter level is used to decide which word should be given as a response. The main problem with the model in the current context is that it is deterministic; that is, for a specific stimulus word, the model always produces the same activation value. The model can produce errors, but only by placing variability in the decision rule: The probability of a particular response is based on the relative amounts of activation across all possible responses (Luce's choice rule; Luce, 1959).

Changing the model so that it would not be deterministic, so that the output activation value for a word would not always be the same on every trial, would require many assumptions about where and how to introduce variance, what the stopping rule for processing should be, and how to map activation values against decision criteria. Proposals have been made (Cohen, Dunbar, & McClelland, 1990; Jacobs & Grainger, 1992) that add assumptions to existing models to attempt to overcome some of the problems, but these proposals are limited to their specific domains.

The Seidenberg and McClelland (1989) Model

Like the interactive activation model, the Seidenberg and McClelland model was designed to explain decisions about words but, unlike that model, it represents information across nodes in a distributed fashion. One level of nodes represents orthographic information, another phonological information, and a third is a hidden layer of nodes between the other two. Orthographic information about a word stimulus is input to the orthographic layer and then activation flows among the nodes of the layers in a single iteration and the output is computed. For a naming response, the output at the phonological layer is compared with the correct representation for each word, and the best matching word is chosen. For lexical (word/nonword) decisions, the output pattern of activation at the orthographic layer is matched against the representation that was input to the system from the stimulus. A good match indicates a word decision, a bad match a nonword decision. The assumption about how decisions relate to response times is that response time depends on the quality of the match—a poorer match leads to longer decision times. However, this assumption is only a promissory note for the model because it is not obvious how it could support predictions about the full range of time-dependent variables.

The Matched Filter Model (Anderson, 1973)

One of the earliest neural network models designed to deal with psychological data was the matched filter model. It was developed to explain performance in the fixed-set Sternberg paradigm (Sternberg, 1966, 1969). In the relevant version of this paradigm, a small set of items is designated to receive a positive response and another set of items is designated to receive a negative response. The model assumes that items are represented as vectors of features, and learning is the formation of a positive filter by summing the vectors of the positive items and the formation of a negative filter by summing the vectors of the negative items. The vector for a test item is compared with these two filters by taking the dot products of the vector with each of the filters. The output of this comparison process is accumulated over time until it reaches either a positive or a negative criterion. Although the model can accurately predict mean response time as a function of the size of the positive set and it can predict some sequential effects, it was not designed to account for accuracy or for the shapes of reaction-time distributions. Anderson's (1991) BSB model can be seen as an update of the matched filter model that does attempt to account for a fuller range of measures.

Summary

These early connectionist models were designed to address questions about structure and process in various cognitive tasks. For the models that deal with reaction time as a dependent variable, the primary concern was the behavior of mean response time across experimental conditions, and the models were not concerned with or able to account for detailed characteristics of the distributions of response times or relationships between reaction time and accuracy. These models were impressive in how much they accomplished, but newer models have been designed to take the next step to the full range of phenomena associated with reaction time and accuracy.

Models Based on the GRAIN Framework

The GRAIN framework (McClelland, 1993) was designed to guide the construction of models in which decisions are based on interactive processes that evolve over time, with variability built into processing. The goal is to account for response time and response probability for the decisions required by a task. As in earlier connectionist models, processing in GRAIN models is assumed to take place in a continuous manner as information or activation flows gradually and interactively through a connectionist network. The key feature that makes GRAIN models different from earlier models is the introduction of variability into processing. Each time activation is input to a node, the amount that is input includes a random number. This addition of random noise guarantees that both the time to reach a decision and the choice of response will vary across stimuli and across repeated instances of a single stimulus.

The GRAIN framework was presented in two articles. McClelland (1993) provided a set of general principles that define the GRAIN approach, and Movellan and McClelland (1993; see also 1991) applied a GRAIN-based architecture to the problem of learning to produce variable outputs that corresponded to random variables from probability distributions. Because GRAIN provides a general modeling framework instead of a single specific model, we were faced with multiple options in developing GRAIN-based models to test against the data from Experiment 1. In the sections that follow, we explain our choices.

In GRAIN-based models, activation flows through a network in a series of cycles. For a three-layer model, activation flows among input, output, and hidden layers. On the first cycle for a stimulus, activation from the stimulus is input to each node in the input layer and transmitted to the nodes of the hidden layer. The net amount of activation input to a node is the sum of the activation values of nodes connected to it weighted by their connection strengths, with noise added and the sum transformed nonlinearly to produce activation values between -1 and $+1$. From the hidden layer, activation is transmitted to the output layer. The activation from the stimulus is maintained in the input layer over subsequent cycles, in each of which activation from both the input layer and the output layer is transmitted to the hidden layer, and activation from the hidden layer is transmitted back to the output layer. Each cycle constitutes one iteration or time step. Cycles continue until a criterion level of activation is reached at the output layer. Response time is determined by the number of cycles to reach criterion.

Learning During the Experiment Versus Learning Prior to the Experiment

To perform the signal detection task, a connectionist network has to be trained to make a numerosity judgment; that is, it has to be trained to discriminate between a large number of asterisks and a small number of asterisks. We tried two training methods for GRAIN-based models; training could take place over the course of the trials in a simulated experiment or it could take place prior to the beginning of the experiment. We tested both alternatives. Movellan and McClelland (1991, 1993) examined the ability of a network to learn to produce output values that approximate various probability distributions (such as normal distributions, or binary distributions such as exclusive–or). The number of learning trials they used was typically a few hundred (with a three-layer model), about the number of trials in one session of our experiment, so for the first GRAIN model we tested, we followed Movellan and McClelland (1991, 1993) and had it learn during the trials of a simulated experiment. We used a similar architecture, a similar error-correcting learning rule, and about the same rate of learning as did Movellan and McClelland (1993).

The alternative training method, pretraining, is suggested by McClelland, McNaughton, and O'Reilly (1995, p. 435; see also Ratcliff, 1990, p. 306), who argued that learning the structure of a domain (such as numerosity) must occur over an extended time period during which the system is exposed to all possible stimuli multiple times and in random order. By this argument, performance in the signal detection task would be a function of long-term knowledge of numerosity. So, in our second GRAIN model, the network was trained to produce, for each possible number of asterisks, the probability with which that number was drawn from the high distribution. This training occurred prior to the tests of the model in simulated trials of the experiment.

Number of Layers

Another choice was between a two-layer model and a three-layer model; both have been used in the GRAIN framework (Movellan & McClelland, 1993; Usher & McClelland, 1995). A two-layer model would be sufficient for data from the signal detection task, but for our main investigations, we chose to implement three-layer models. A three-layer model is the model most often used for tasks such as word identification, lexical decision, and word naming (McClelland & Rumelhart, 1981; Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989), and we wanted our conclusions to generalize to those tasks. We also wanted it to be possible for the model to apply to other tasks that use the asterisk stimuli, for example, a same–different matching task in which two displays are presented and the subject is required to decide whether both have a high or low number of asterisks (respond "same") or one display has a high number and the other a low number (respond "different"). Ratcliff and Rouder (1998) have examined a task of this kind using stimuli varying in brightness. A two-layer model would not be able to perform this task because it is logically equivalent to the exclusive–or task that two-layer models cannot perform (see Movellan & McClelland, 1993, and discussion in Plaut et al., 1996).

Learning Rule

To perform the signal detection task, the network must be trained to respond high to a high stimulus and low to a low stimulus. Although the GRAIN framework does not specify learning rules, we followed Movellan and McClelland (1993) in using a version of contrastive Hebbian learning algorithm (which is similar to the mean field learning algorithm; Peterson & Hartman, 1989; see also the application by McCloskey & Lindemann, 1992). This is an iterative error correcting rule that behaves in a manner similar to the noniterative backpropagation rule.

Net Input Averaging Versus Activation Averaging

The net input to a node is sometimes calculated as the running average (i.e., a weighted sum) of the prior averaged net input with the current net input (e.g., McClelland, 1993) and sometimes as the running average of the amount of activation at a node (McClelland & Rumelhart, 1981). Both methods are used to smooth out large fluctuations in amount of activation. We focus on the former but report results for both.

A GRAIN-Based Model With Learning During the Experiment

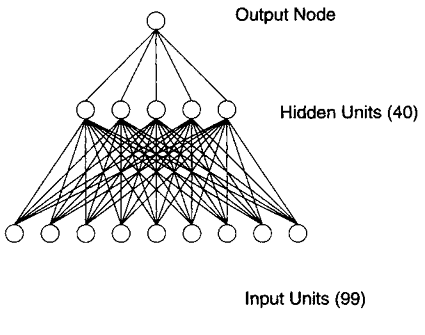
All of the choices about architecture and learning just described determined the several GRAIN models to be evaluated using the signal detection data. The model that appeared to us at the beginning of this project to be most plausible and most likely to fit the data while still allowing generalization to other cognitive tasks was a three-layer model, with learning taking place during the experiment rather than prior to it. We assumed that subjects come to the experiment having learned to represent numerosity, but they must learn the experimenter-defined high and low response probabilities that are appropriate to the task. We discuss this model first.

Given the architecture and the form of training to be used, there were still various ways the stimulus information—the numbers of asterisks—could be represented. The scheme we chose was to represent a stimulus composed of N asterisks as a vector of length 100 (to allow numbers up to 100) with element N set to 1, and elements around N set to 1, and all other elements set to 0. The elements around N were set to 1 to allow generalization across similar numbers, and the number of such elements, called *window size,* was a parameter of the model.

The three layers of the model were an input layer of 100 nodes to represent a stimulus input vector, an output layer of one node to indicate a response (as in Movellan & McClelland, 1991, 1993), and a hidden layer of 40 nodes, with each hidden layer node connected to the output node and to each input node. The initial weights on the connections between nodes were set to random numbers from a uniform distribution between +1 and −1 (the size of this range is a parameter of the model). The activation value in the single output node determined a response: high if the node's activation value was near +1 or low if it was near −1. Note that these response criteria have a function similar to the boundaries for high and low responses in the diffusion model. Figure 16 illustrates this GRAIN model, and a full description of the model and equations is presented in the Appendix.

Our simulations used about the same number of trials as subjects received in a whole experiment (i.e., 12,000). As a check on the accuracy of the simulations, the first two authors of this article implemented the model independently. Stimuli were input to the model over trials sequentially, each stimulus with feedback, just as for the subjects of Experiment 1. Parameter adjustment to achieve the best possible fits of the model to the data was accomplished by hand because there was insufficient computer power to embed the model in a minimization routine (see the discussion at the end of this section).

With the contrastive Hebbian (mean field) algorithm, there are two phases, a free phase during which activation from the stimulus is held constant at the input nodes and a clamped phase during which activation is held constant at the input nodes and the desired output is held constant at the output node (for details, see the Appendix). To simulate the decision process for a stimulus, a vector corresponding to the stimulus was input to the network at the input level, and activation was allowed to flow from the input and output layers to the hidden layer and from the hidden layer to the output layer. In the initial version of the model, the input to each node at each time step was the running average of net input. Free phase cycles continued until the value in the output node reached either the positive (near +1, high) or negative (near −1, low) criterion. The number of cycles was used to represent response time. Then feedback was presented to the system by entering 1 or −1 in the output node, and activation was allowed to flow with the output node clamped at the feedback value (and the input nodes still at their original stimulus values). The contrastive Hebbian algorithm was used to modify the weights connecting



Input Representation

000000000000111111111000000000......

For the number 17 with a window of size 9.

Figure 16. An illustration of the architecture of the first GRAIN-based model and the input representation.

each pair of nodes, with the amount of modification depending on the difference between the product of the activation values in the nodes connected by the weight after the free phase and after the clamped phase, multiplied by a learning rate parameter (see the Appendix).

There were eight free parameters for the model (plus some parameters to scale number of iterations to time). One was the window size, the number of positions around the stimulus value that were set to 1. Another was the absolute value of the criterion for a response from the output node, a positive value for a high response and a negative value for a low response. There were also the learning rate parameter ϵ for the contrastive Hebbian algorithm and the parameter m, the absolute value of the limits of a uniform distribution from which random numbers for the initial weights were chosen. There were also two parameters that determined the flow of activation, λ and σ. To obtain the running average of net input at a node, a proportion (λ) of the node's prior average input was added to $1 - \lambda$ of the node's current input. σ was the standard deviation in the value of noise (mean = 0) added to the net input. Finally, there were two parameters that controlled the iterative process (see the Appendix).

Many simulations were run to check different sets of parameter values, and the fits of the model to data that are presented in the sections below come from the most successful set of parameters, which are shown in Table 3. The fits of the model are presented for two sets of response criteria values (± .90 and ± .85), to show whether criteria manipulations might account for the behaviors of different subjects.

Learning

The model was fit only to asymptotic data; the first 1,000 trials were discarded. It took several hundred trials before the simulated data became relatively stable (e.g., high responses were slower for the first 400 trials than the asymptotic value, although low responses asymptoted after 200 trials in one simulation). Because extreme values (e.g., 10 and 90 asterisks) did not occur very often, it took more than 300 trials before enough were presented for the simulation to accurately classify them.

Table 3
Parameters of the GRAIN Model

	Parameter values		
Parameter name	Averaging net input	Averaging activation	Fixed weight model
Learning rate, η	.2	.4	0
Window size in the input representation	±8	±8	±8
Initial weight range (±m)	1.0	1.0	.5
Running average of the net input, λ	.10	.20	.35
Response criteria	±.90 (±.85)	±.90 (±.95)	±.75 (±.80)
Standard deviation in the noise in the net input, σ	.25	10.00	1.20
Value of τ in the annealing schedule	10.	2.	2.

Note. The first column of parameter values is used in the body of the article, the second column of parameter values is used in the Appendix.

Response Probability and Sequential Effects

Figure 17 shows the model's predicted probabilities of low responses across all of the experimental conditions (all of the possible numbers of asterisks), for four sequential conditions. The top panel shows the functions with the criteria for responding set at a value of activation in the output node of ± .90, and the bottom panel shows the same functions with the criteria ± .85. The general shape of the functions is the same as for the data (Figure 1). However, the predicted sequential effects are larger than in the data, and more important, they are dependent only on prior feedback, whereas the data show sequential effects dependent only on the prior response. There is no way to change this behavior; the model has to predict sequential effects based on prior feedback because it is the feedback that controls weight changes (learning) in the network across trials.

Response Probability and Mean Reaction Time

Figure 18 shows the model's predictions for response time as a function of experimental condition. The predictions follow the data (Figure 2) for correct responses, in that response time slows as the number of asterisks nears the point at which the high and low distributions cross. However, in the data, error responses speed up as errors become more extreme, whereas the model predicts that error responses slow down as errors become extreme. This misprediction is indicated by the fact that, in Figure 18, the response-time functions for all types of responses are monotonically increasing from high-response probability to low-response probability.

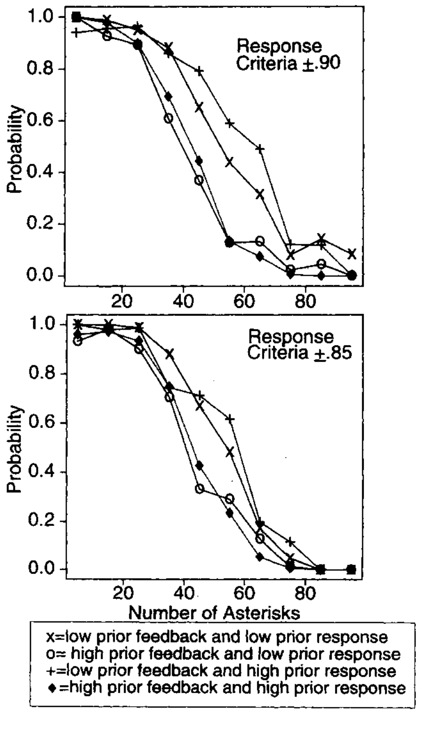


Figure 17. Accuracy functions predicted from the GRAIN-based model with learning during the experiment for the data from Experiment 1 (see Figure 1).

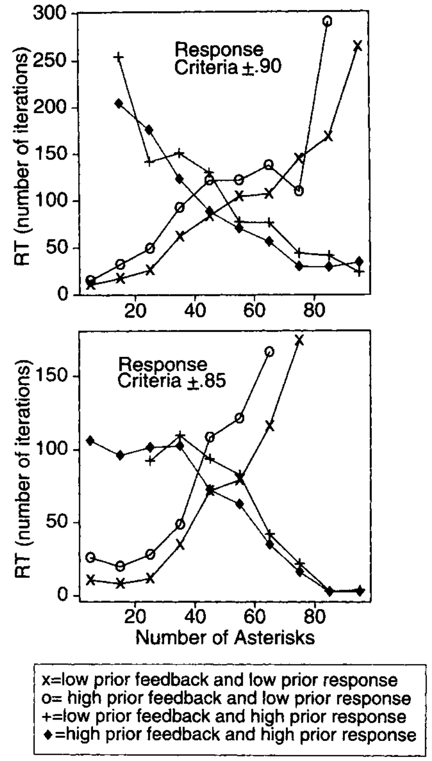


Figure 18. Reaction-time (RT) functions predicted from the GRAIN-based model with learning during the experiment for the data from Experiment 1 (see Figure 2).

The conversion to latency–response-probability functions is shown in Figure 19. Generally, the model predicts monotonically increasing functions from correct to error responses, whereas the data show inverted U-shaped functions. No parameter manipulations affected this pattern of results.

Reaction-Time Distributions

Figure 20 shows predicted reaction-time distributions for several levels of response probability. The predicted distributions are skewed to the right, which matches the data. The extent of the skew, as measured by the distance between the distribution tail and the mode, shows a reasonable approximation to the experimental data.

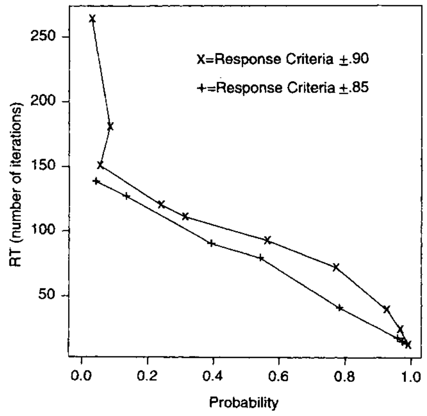


Figure 19. Latency–response probability functions predicted from the GRAIN-based model with learning during the experiment for the data from Experiment 1 (see Figure 3).

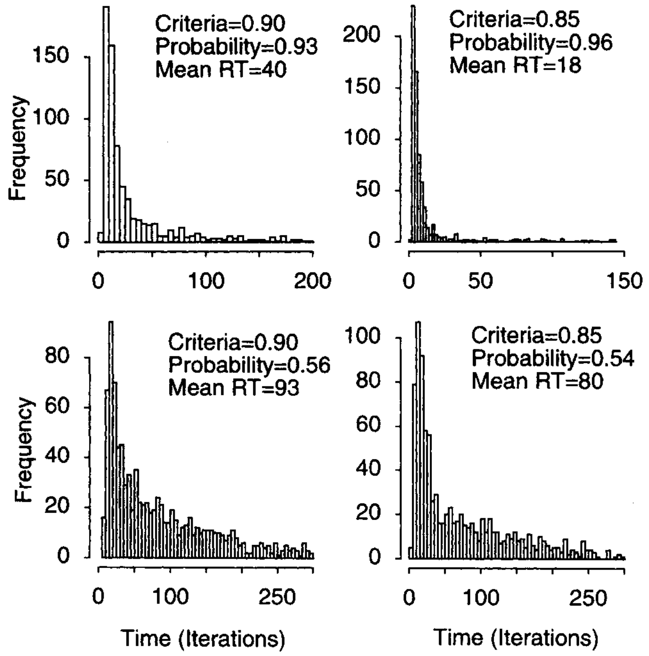


Figure 20. Reaction-time (RT) distributions predicted from the GRAIN-based model with learning during the experiment for the data from Experiment 1 (see Figure 4).

Figure 21 shows the predicted hazard functions for the distributions shown in Figure 20. Generally, they show a rapid rise to a peak followed by a slight fall, like the data (Figure 6), but then followed by a rise in the tail, unlike the data. For the two distributions with response probability around .95, the rise in the tail is in the extreme tail of the distribution where there is numerical instability in the estimate of the hazard function. For the distributions in Figure 21 with response probability around 0.5, the rise in the tail occurs right after the mean reaction time for that condition. In this case, the rise in the hazard function occurs in a region that is not in the extreme tail of the distribution and so cannot be attributed to instability (see Glaser, 1980, for a discussion of such bathtub-shaped hazard functions). Thus, although the reaction-time distributions look plausible, the hazard functions in the low-probability conditions differ significantly from the data in the portion of the distribution beyond the mean.

Summary

Although this model captures some features of the data, it fails in several important respects. It incorrectly predicts sequential effects based on prior feedback instead of on prior responses; this incorrect prediction comes from the use of the learning rule throughout the sequence of simulated experimental trials. The model also mistakenly predicts that error responses are always slower than correct responses, and so it fails to capture the data's

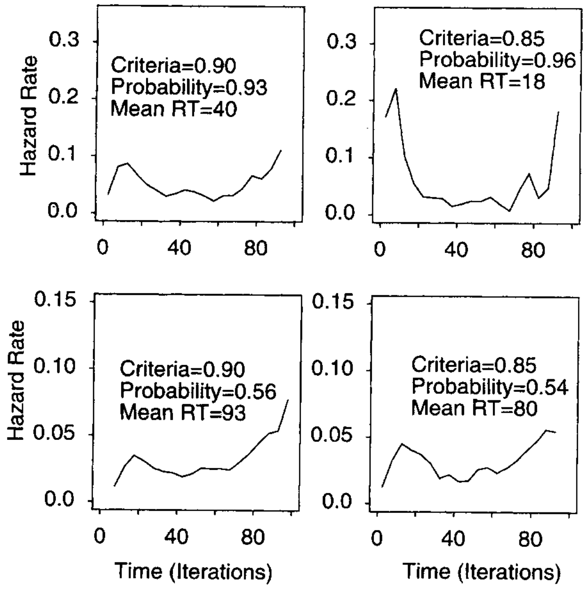


Figure 21. Reaction-time (RT) hazard functions predicted from the GRAIN-based model with learning during the experiment for the data from Experiment 1 (see Figure 6).

inverted U-shaped latency–probability functions. It also fails to predict the correct shape of the hazard functions.

Parameter values. Given that this model could not correctly account for sequential effects or latency–probability functions, it is of concern whether other parameter values might have led to more success. The fits of the model that were presented above are the best that we could find. Altering the parameter values led to the following problems (all comparisons are described relative to the results presented above):

(1) Learning rate, ϵ, in weight modification in the contrastive Hebbian algorithm. When the learning rate is reduced, there are fewer extreme errors and there are still large sequential effects in the data. When the learning rate is increased (e.g., 0.2 to 0.4), then the extreme stimuli (0–10 and 80–100) have fast reaction times with no variability (i.e., there is no distribution of reaction times, just a single value for all responses). To examine what happens when learning is turned off after initial training, we used a learning rate of $\epsilon = 0.4$ for the first 4,000 trials, then reduced it gradually to $\epsilon = .001$ over 600 trials, and then kept it at that value from then on. For trials after the learning rate had been reduced, the size of the sequential effects was reduced (to about .05 in response probability), but extreme stimuli had no variability in response time (the number of steps to criterion was 11 and there were no errors for stimuli 1–30 and 80–96, whereas response time is in the range of 100–400 steps for stimuli that produced errors). Errors were also slow relative to correct responses (e.g., 100 to 350 steps slower for the more extreme errors).

(2) Window size in the stimulus representation. When this is made small (e.g., ± 4 instead of ± 8), all processes terminate in the minimum number of steps and there is no variability in reaction time. Also, response probability is at chance over much of the range from 20 to 80 asterisks. When the window size is increased (e.g., to ± 12), results are much the same as for window size ± 8.

(3) Initial weights. When the initial weights are increased from a range of $-.5$ to $+.5$ to a range of -5 to $+5$, the behavior of the model does not change significantly.

(4) Running average parameter (λ). Changing the size of this parameter from .05 to .75 alters the minimum reaction time but does not significantly affect the qualitative behavior of the model with respect to reaction times or response probabilities and does not change sequential effects.

(5) Response criterion. When the response criterion is decreased from .9 to .7, then most processes finish in 1, 2, or 3 iterations. When the response criterion is raised above .95, reaction time is slowed a little, but the qualitative behavior of the model is not altered.

(6) Annealing schedule. In the iterative decision process, the transformation from net input to activation is adjusted to make a fixed net input have a larger effect on activation after each iteration (this is called *simulated annealing*). In the formula mapping net input to activation (see the Appendix), net input is divided by a constant. In the contrastive Hebbian algorithm, the value of this constant (τ) is reduced on each successive iteration to make the same size net input have a larger effect on activation. In practice, this is done by multiplying τ by .99 on each iteration (Peterson & Hartman, 1989).

Eliminating annealing makes a large difference to reaction times. Responses in extreme conditions (high and low numbers of asterisks) are made in four iterations, whereas in nonextreme conditions most processes do not terminate in 2,600 iterations. When the parameter τ is reduced to 1 from 10 (see Table 3), responses are made in about 2 or 3 iterations in the extreme conditions, and in the less extreme conditions, many processes do not terminate in 2,600 iterations. This produces distributions with tails that are far too long relative to the mean.

Other possibilities. Later, we investigate a GRAIN-based model for which learning takes place prior to the experimental trials. Here, we discuss various alternatives for the model with learning during the experiment. Some of these might appear to be minor alterations, but they can have a large impact on the predictions of the model.

If a running average is computed on activation rather than net input, then the maximum value that activation can be changed as a result of a very large value of noise is λ, the proportion of new input to be added to the running average activation value. When the running average is over net input, a large value of noise can change activation by any amount from the prior value to ± 1. This means that changes in activation will be more gradual when the running average is over activation than when it is over net input.

Many predictions of the model with a running average over activation were similar to predictions with running average over net input. The accuracy functions were similar, and sequential effects still followed prior feedback rather than the prior response. Latency–probability functions still increased from correct to error responses, but in a few conditions, the most extreme errors were fast (reflecting fewer than five responses out of thousands in the simulation). These latency–probability functions were not at all similar to the inverted U-shaped functions exhibited by the data from Experiment 1. Also, the reaction-time distributions had extremely long tails and a small proportion of processes did not terminate in the 2,600 iteration limit in the computer program (or, in some tests, by 5,000 iterations). The means of the distributions were in the tails of the distributions, well beyond the mode, which is unlike the experimental data in which the mean occurred only a little later in the tail than the mode. The corresponding hazard functions were highly peaked and fell to asymptotes that were less than one fourth their peak, in contrast to the data in which the asymptotes were only a little below the peak. The parameters for this version of the model are shown in the second column of Table 3.

We also considered two other ways of representing the number of displayed asterisks. For one scheme, N (the number of displayed asterisks) of the elements of the 100-element input vector were randomly chosen to be assigned the value +1 and the other elements were assigned the value −1. The second scheme used a thermometer representation in which the number of asterisks displayed was represented as +1 in the first N elements starting from 0. For example, for the number 30, the elements 1–30 were assigned +1 and the nodes from 31–99 were −1.

The problem with both of these schemes was that response probability failed to reach ceiling or floor as the number of asterisks became extreme (e.g., below 30 or above 70) and the S-shaped response-probability functions had a much lower slope than in the data. For the random assignment representation, the source of the problem is likely the fact that randomly assigning inputs to the input nodes (turning them on or off depending on the random assignment) produces no consistent mapping from the number of asterisks in the stimulus to an input representation that can be used in learning. For the thermometer representation, the low units are given inconsistent training; they are trained to a low response for low stimuli but they are trained to a high response for high stimuli. For both schemes, we manipulated learning rate, the proportion of old net input averaged with the new net input, the amount of variability added to activation, and the annealing–scaling parameter (τ) in the nonlinear transformation from net input to activation, but none of these manipulations altered the results.

Another possibility was that the model's predictions could be improved by taking into account the variability with which subjects encoded the stimuli. To test this, we used the window representation, and we assumed that for a stimulus number N, the number actually encoded was normally distributed around N with mean 0 and standard deviation 8. This had almost no effect on the model's predictions. In effect, adding this variability is computationally equivalent to increasing the standard deviation in the high and low distributions by a factor of 20% or so.

There may be other ways to produce sequential effects more consistent with the data. For example, it might be assumed that response alternatives are primed by the prior response. It is unlikely that this mechanism would work because the model predicts large sequential effects based on the prior feedback (larger than the effects observed in the data) and these would be added to the effects produced by response priming.

We also examined a two-layer version of the model. We were mainly interested in the behavior of three-layer models because three layers are required to perform interesting behavioral tasks (e.g., McClelland & Rumelhart, 1981; Seidenberg & McClelland, 1989). However, for completeness we note results for a two-layer model. In the two-layer model, each input node was connected to the output node directly. The simulations produced results very similar to the results for the three-layer model, both with net input averaging and with activation averaging. For example, for the net input averaging version, latency–probability functions were still monotonic, unlike the inverted U-shaped data. The sequential effects in response probability showed large effects of prior feedback just as in the three-layer model. The reaction-time distributions had long tails, and for the distributions with response probability around .5, the hazard functions rose immediately after the mean reaction time, unlike the data in which the functions were roughly constant after the mean reaction time.

In sum, we could find neither alternative parameter values nor alternative assumptions about structure or processing that would improve predictions for the GRAIN-based model with learning during the experiment. In several cases, it seemed that one manipulation might fix one deficiency and another might fix the problem that the first introduced, so that both together would improve predictions. Whenever this appeared possible, we evaluated the manipulations jointly, but there were none that produced results better than those presented above. The model always predicts sequential effects dependent on prior feedback.

The main reason that the model could not provide an account of the data is that the information used by the model to learn the task (feedback on whether the stimulus was selected from the high or low distribution) is representative of only part of the data to be explained. The learning algorithm has no access to any information about what response times should be, and so the network cannot be adjusted during training according to differences between its predicted response times and the response-time data.

A GRAIN-Based Model With Learning Prior to the Experiment

One major problem with the GRAIN model just discussed was that the learning rule led to sequential effects that were dependent on prior feedback. In an attempt to avoid that problem, we examined a GRAIN-based model for which the weights on the connections among the input, hidden, and output nodes were fixed by training prior to simulation of the experiment.

The first question for this model was how to train it to distinguish between high and low numerosity. For the diffusion model, the underlying variable that controlled performance was the probability p that a stimulus came from the high versus the low distribution. We decided to train the GRAIN-based model using this probability. The model was trained so that an input number of asterisks would give as output a linear transformation of the probability that number of asterisks came from the high distribution. The algorithm was that used for the first model but without variability in activation or net input. This model had the same architecture as the model with learning during the experiment, an input layer of 100 nodes, a hidden layer of 40 nodes, and an output layer of one node (as shown in Figure 16).

In the preexperiment training phase, stimuli were presented to the network in random order. Given some stimulus number of asterisks to the input nodes, the network was trained to produce an output activation that was the stimulus probability $0 < p < 1$, transformed to the range −.8 to +.8 (i.e., $1.6p - .8$). The feedback provided at the output node was the probability with which the stimulus came from the high distribution, also scaled to lie between −.8 to +.8. This scaling was necessary because training the network to produce activation values of −1 and +1 to extreme stimuli produced ceiling and floor effects in reaction time; for example, for stimuli with 1–20 or 80–96 asterisks, all processes terminated in exactly the same number of steps. The learning rate for the contrastive Hebbian algorithm was 0.1, the running average parameter was 0.80, the annealing–scaling parameter was zero, the initial weights varied from −0.5 to +0.5, there were 2,590 iterations unless the output activation was within .001 of the target activation, and no noise was added into the system. With these parameters, sufficient training trials were run (30,000, but this was not manipulated) so that the weights produced (transformed) probability values at the output node within about .05 of the target values.

After training, the behavior of the model was evaluated on experimental trials presented to the network just as they had been for human subjects. Variability (noise) was added to net input before transformation to activation. The parameters of the model (except the connection weights that were fixed by prior training) were adjusted by hand until the model produced results that qualitatively matched the data.

Response times for errors were about the same as or faster than response times for correct responses, like the experimental data for Subject 4 but unlike those for Subjects 1, 2, and 3. This means that the model failed to produce the asymmetric inverted U-shaped functions typical of the data.

To attempt to produce the experimental pattern of data obtained for Subjects 1, 2, and 3 (slow error reaction times relative to correct responses), we added a new source of variability into the model, the equivalent of drift variability, derived from our experience with the diffusion model. Specifically, we assumed that the same stimulus was not always encoded the same way across trials, so that given some number of asterisks as a stimulus, the model was presented not with that number, but with that number plus a number drawn from a random normal distribution with mean 0 and standard deviation $r = 8$. For example, the stimulus of 40 asterisks might be input to the model as 32 asterisks, 48 asterisks, or any number in between. This variance mimics the variation across instances of a stimulus that is part of the diffusion model (the parameter η).

Overall, with the addition of variable encoding, the model does a good job of accounting for reaction times for correct and error responses, the probabilities of those responses, and with minor exceptions, the shapes of reaction-time distributions and hazard functions. However, like the earlier model, the model does not account for sequential effects. The parameters of the model we used to obtain the qualitative fits are shown in the third column of Table 3.

Response Probability, Reaction Time, and Reaction-Time Distributions

Response probability as a function of number of asterisks mimics the human data (Figure 22). Mean reaction times slowed as the number of asterisks neared the crossover point (Figure 23) as in the human data, extreme errors were faster than correct responses, and errors for conditions closer to the crossover point were slower than correct responses. The latency–probability functions from the model are quite similar to those for Subjects 1, 2, and 3 (Figure 24).

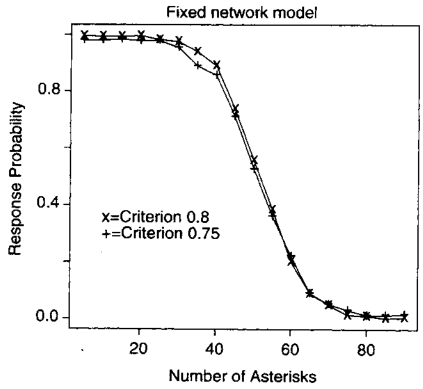


Figure 22. Accuracy functions predicted from the GRAIN-based model with learning prior to the experiment for the data from Experiment 1 (see Figure 1).

Reaction-time distributions (Figure 25) are skewed to the right just like the human subject data (Figure 3). However, the fastest responses have the same number of iterations across conditions and do not slow as mean reaction time increases, as they do in the experimental data (e.g., a 20-ms effect for Subject 3). The hazard functions (Figure 26) appear to increase and level off for the high-probability conditions (before variability in the tail of the distribution makes the estimates unstable). However, the hazard functions with response probability around .50 (bottom panels in

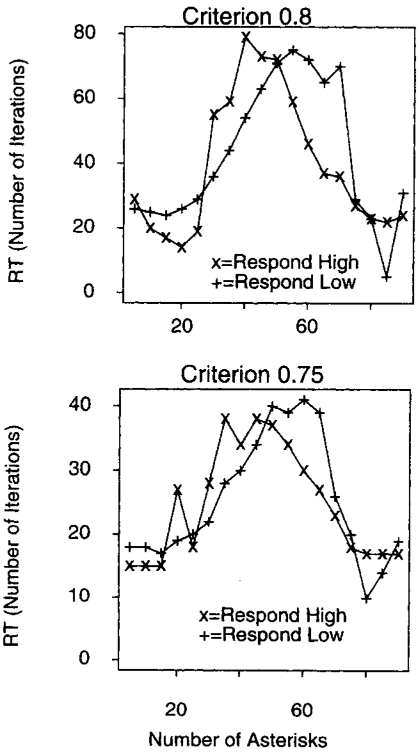


Figure 23. Reaction-time (RT) functions predicted from the GRAIN-based model with learning prior to the experiment for the data from Experiment 1 (see Figure 2).

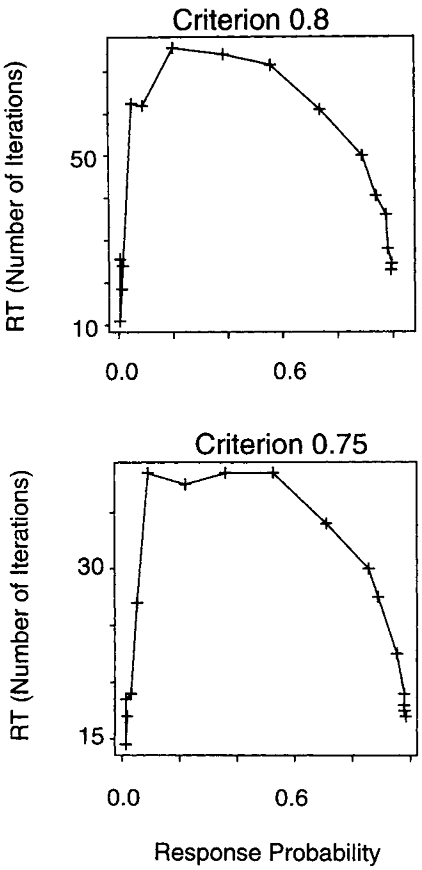


Figure 24. Latency–response probability functions predicted from the GRAIN-based model with learning prior to the experiment for the data from Experiment 1 (see Figure 3).

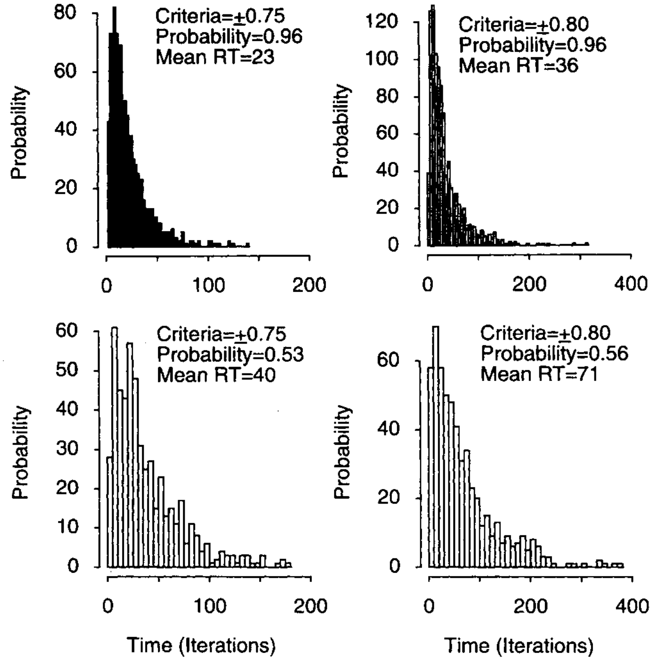


Figure 25. Reaction-time (RT) distributions predicted from the GRAIN-based model with learning prior to the experiment for the data from Experiment 1 (see Figure 4).

Figure 26) rise rapidly immediately after the mean reaction time, which does not match the experimental data.

Sequential Effects

The model does not produce any sequential effects, and there is no single assumption that could be added to produce the different patterns of sequential effects for the subjects in Experiment 1. The problem is that different subjects produced different patterns of sequential effects; one showed no effect of prior response, one produced the opposite response with greater probability, and two produced the same response with greater probability. Any assumption about changing criteria or residual activation from one trial to another would have to be different for each subject.

Discussion

Except for sequential effects and some details of reaction-time distributions, this model provided a reasonably good qualitative explanation of the data from Experiment 1. However, this would not have been possible if we had not had the knowledge we gained from the diffusion model fits, namely that variability in encoding of the stimuli across trials was needed to produce slow errors and that stimulus probability was an appropriate training function. Without across-trial variability, the model could not have correctly accounted for the relations between correct and error response times. Without prior knowledge of what function to use for training, we would have had to resort to guesswork (and some guesswork was needed because a linear transformation of stimulus probability was required). Of course, even though stimulus prob-

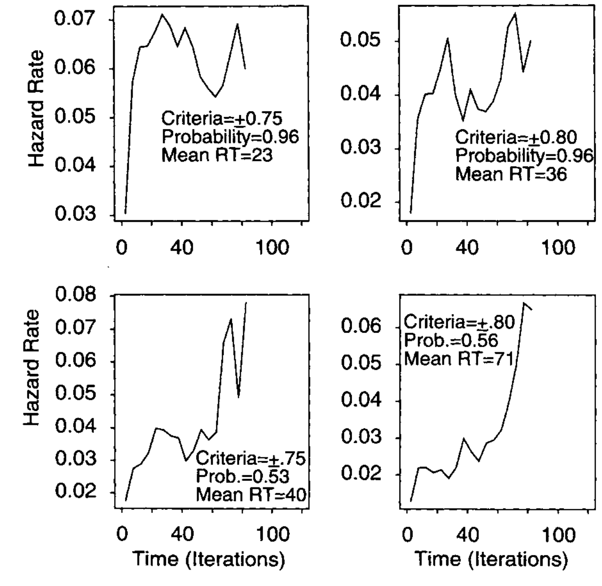


Figure 26. Reaction-time (RT) hazard functions predicted from the GRAIN-based model with learning prior to the experiment for the data from Experiment 1 (see Figure 6).

ability was not intuitively obvious to us as a candidate training function, pointers might have been gained from earlier literature such as the probability matching literature (e.g., Estes, 1995). But intuitions about training functions cannot extend to all possible situations. If payoffs or instructions were used to produce a bias toward one response (see Ratcliff & Hacker, 1981), then stimulus probability might not be the correct training function. Also, finding a training function would become difficult if the structure of the problem was not linear (e.g., two criteria or nonlinear boundaries in a 2-D categorization space; Maddox & Ashby, 1993), and finding a training function would become very difficult if subjects did not discover the true statistical structure of the problem and so responded according to some function of their own invention.

If intuition cannot reliably provide a training function for a connectionist model, then the question is whether it is possible to develop a method or computer program to search for and find an adequate training function. The functional form of a training function would have to be assumed (e.g., for the asterisk task a logistic function might be assumed), and the program would have to search for parameter values for that function that would allow the model to fit the data. However, this issue is complicated and requires some discussion.

First, some idea of whether the model could produce good qualitative fits of both response time and accuracy data would be required, independent of the training function. A model that could not, for example, produce errors faster than correct responses might distort fits to correct response times in an effort to produce faster errors. In this case, poor fits of the model to the data would have nothing to say about the training function.

If it seemed that the model could qualitatively fit the data, then the model would be trained to asymptote (without variability in activation values) to reproduce the training function (as in our second GRAIN model). Then it would produce predictions for the dependent variables (with variability in activation values so as to produce distributions of reaction time and errors), and then the parameters of the training function and the other parameters of the model would be adjusted on the basis of the discrepancy between the predictions and the data; then the cycle of training, testing, and adjustment would be repeated. This multiple recycling of training, testing, and parameter adjustment would be difficult enough, but the difficulty would be made even greater by the large number of parameters that would be required: (a) for the learning phase, all of the parameters of the training function plus the learning rate, the initial weights, the learning criterion, and the parameter of the logistic net input to activation transformation and (b) for the test phase, the logistic net input to activation, whether there is a bias unit or not, the running average parameter, the annealing parameter, the amount of noise, the response criteria, variability in mapping from the stimulus to the input value, the mapping from number of iterations to reaction time, and the nondecisional component of reaction time.

To be specific, for Experiment 1, there are 96 stimulus values. One approach would be to assign a different parameter to represent the target value for training for each stimulus value. This would result in an unmanageable fitting program (fitting more than 100 parameters could take years). Reducing the number of stimulus target values to 10, each spanning a 10-digit range from 1–96 (or assuming a simple three-parameter training function), might be a way of producing a faster program. Even with some reduction in the number of parameters, there would be practical issues of speed. For the diffusion model, on a fast workstation, a single set of predictions can be generated in about 1 min. An optimal set of the six or seven parameter values that produce good fits to the data can be obtained in a few hours (with several hundred iterations). In contrast, for the connectionist models examined here, it takes an hour or two to train the network and then produce one set of predictions. A program to perform a series of training and prediction trials and then adjust parameters on the basis of reaction time and accuracy would take weeks or even months to produce one set of optimized fits for one set of data (even if the number of parameters were reduced to 10 stimulus values plus the other parameters of the model).

Another complicating factor for many connectionist models is that they have distributed representations. This means that they have to be trained on all stimuli at once, because for some learning algorithms, learning stimuli individually would result in catastrophic interference (McCloskey & Cohen, 1989; Ratcliff, 1990). With algorithms that do not suffer from catastrophic interference (e.g., the BSB model) but have forgetting built in, early training would be forgotten after later training. The diffusion model does not have this kind of problem because it provides a way of determining the drift rate for each stimulus independent of the others, and relationships across stimuli are assessed after the model has been fitted to obtain the separate drift rates.

Setting aside the problems of developing algorithms to automatically find the parameters of an assumed training function for a connectionist model, there is still the problem of the plausibility of the basic notion of pretrained networks (i.e., how would humans represent this information and use it). Implicit in the search for an appropriate pretraining function is the idea that people perform tasks such as the asterisk signal detection task with preexisting networks, assemblages of networks, or parts of networks (e.g., Usher & McClelland, 1995). Some plausibility is given to this idea when subjects already know a lot about whatever is relevant to an experimental task before they come into an experiment and when they learn a task very rapidly. But the problem is how, for any given task, a network with the appropriate properties is chosen or assembled from all of what would have to be a multitude of preexisting networks. For example, long-term knowledge about numerosity can be used for many different tasks, and tasks with the same statistical structure can be performed with many different kinds and dimensions of stimuli. To assemble a network for one particular task, many different issues would have to be considered: what dimension is involved (e.g., numerosity, tone frequency, word familiarity), what scale is involved in the task (e.g., for numerosity, 1–10 or 1–1,000), what is the function relating stimuli to response choices, how are "signal" and "noise" represented, and so on. At the present time, connectionist modeling with pretrained networks has not begun to address these issues.

In conclusion, although the pretrained connectionist model does a reasonable job of accounting for the experimental data, it leaves important larger questions unresolved. Neither the model nor the approach embodied in the GRAIN modeling framework gives any account of how to assemble networks for specific tasks or how a learning function for pretraining networks might be chosen.

Anderson's BSB Model

Anderson (1991) has applied the BSB (Anderson, Silverstein, Ritz, & Jones, 1977) to explain reaction times in same–different letter string matching paradigms. The BSB model assumes that a stimulus item is represented by a vector of elements (called the *state vector*), and that memory (prior experience) is represented by a matrix of elements. Memory for a single item is the matrix composed of the products of each pair of elements in its vector (the product of the vector and its transpose). Memory for all items is a single matrix that is the sum of the matrices for each of the items. When a test item is input to the system, its vector is multiplied by the memory matrix and a vector is produced as output. If the test item was previously learned by the system, then the output vector matches the input vector, with some variability that depends on the other items learned.

For the signal detection paradigm, the vectors for the stimuli were divided into two parts, with 99 elements representing the stimulus and 10 elements representing the response to be learned. "High" stimuli were represented by setting the response elements to +1, and "low" stimuli were represented by setting the response elements to −1. When a test item was input to the system, the 10 response elements of its state vector were set to zero and the vector was multiplied by the memory matrix to produce output in the response elements. This new state vector was multiplied by the matrix again to produce a better representation in the response part of the vector, and this process was iterated until the representation of the response reached some criterion. The number of iterations was taken to be the reaction time for the response.

Following Anderson (1991), the iterative vector–matrix multiplication process was augmented by two other factors to produce stability over iterations. First, each new state vector for the next iteration was the product of the vector–matrix multiplication plus some proportion of the original input vector (so the input part of the vector would not change too much across iterations) and some proportion of the previous state vector. The updating rule used was

$$x(t + 1) = \gamma x(t) + \alpha A x(t) + \delta f(0),$$

where $f(0)$ is the initial input, γ is a constant a little less than 1, and α and δ are constants. Thus, the vector at any time in the decision process, $t + 1$, was a weighted sum of the vector at time t, the original input, and the product of the vector at time t with the memory matrix A. Second, for any element in the output vector (stimulus or response), when its absolute value exceeded some limit, it was replaced by the limit value. This corresponds to the bounding box in the brain-state-in-a-box. The decision process terminated when the elements in the response portion of the vector reached a response criterion or the number of iterations exceeded some maximum.

In simulating the signal detection task, once a response had been produced, feedback was provided. The correct response was entered in the response portion of the vector (+1's for high and −1's for low) with the stimulus in the stimulus portion of the vector. Then the memory matrix was updated using

$$A = 0.995A + \eta \mathbf{f}\mathbf{f}^{\mathrm{T}},$$

where 0.995 is a constant representing memory decay and η is a constant. The matrix represents the sum of the products of each element of the vector with each other element of the vector. So without decay, element (1, 101) would contain the number of times a 1 in element 1 was paired with a 1 in element 101 minus the number of times a 1 in element 1 was paired with $a - 1$ in element 101 (cf. probability matching, Anderson et al., 1977).

The interactive process in BSB is deterministic: For the same stimulus and the same memory matrix, the same output will always be produced. Anderson (1991) argued that this is a positive feature of the model and that variability in performance comes from variability in the sequence of stimulus–feedback pairs presented to the model across trials.

To implement the model for the signal detection task, we used input vectors with the elements 1–99 for the stimulus and elements 100–109 for the response. The response element criteria for termination of the iterative process were the values +.5 and −.5. The stimulus representation was like that used for our GRAIN-based models (Figure 16). First, some number of elements (the window size) around the stimulus number was set to 1 and elements more distant were set to 0. The value of the window-size parameter was ± 15. For stimuli less than 15 or greater than 85, fewer elements were nonzero (e.g., for stimuli less than 15, elements up to and including the stimulus number and elements 15 higher than the stimulus value were nonzero). Then the vector was normalized to have size 31, so that the sum over all nonzero elements equaled 31.

The model was tested with sequences of stimuli just as for the subjects in Experiment 1. Initially the memory matrix is set to zero, and after a few trials, the model begins to make responses (on the first few trials, it does not produce a response). After about 200–300 trials, performance begins to asymptote because of decay in the memory matrix (decay parameter = .995).

The parameters for the fits presented below were $\gamma = 1$, $\alpha = .2$, $\delta = 1$, $\eta = 2$, the decay constant for matrix updates was .995, the limit value on the vector value was .5, response criteria were set at .5, and the nondecisional reaction time (T_{er}) and mapping from cycles to reaction time were set to 0 and 1, respectively. We adjusted the parameters by hand to produce behavior as close as possible to the qualitative trends observed in the data (especially, the response probability functions and the reaction-time distributions).

Response Probability and Sequential Effects

Figure 27 shows the model's predictions for probabilities of low responses across experimental conditions. The functions are very similar to the data, and unlike the first GRAIN-based model, BSB predicts that a response is affected by the prior response and not the prior feedback. Because the BSB model weights new inputs (and hence feedback on the prior trial) less than prior memory in the updating rule, sequential effects are based on what response the system produced last time rather than on what feedback it was given. However, it is unlikely that the model could be modified to predict the different patterns of sequential effects that were observed for different subjects. For the one subject showing a bias away from the prior response, the model would have to weight the opposite of the prior response, which would damage learning.

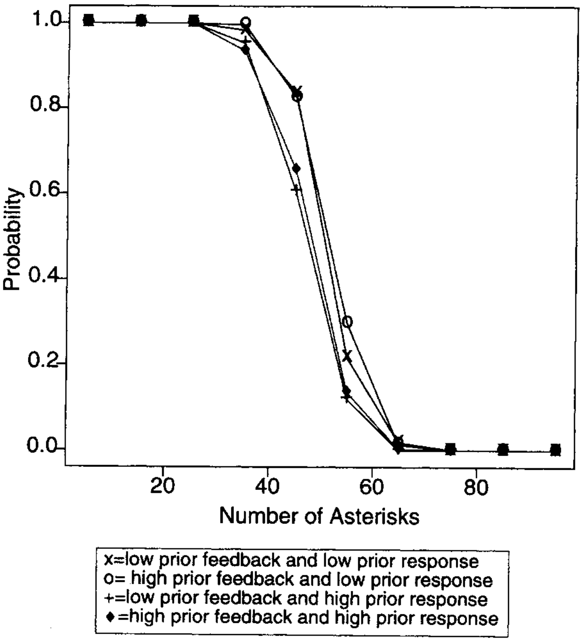


Figure 27. Probability of responding low predicted from the brain-state-in-a-box model for the data from Experiment 1 (see Figure 1).

Response Probability and Mean Reaction Time

The latency–response probability function, shown in Figure 28, diverges from the data because of the model's inability to produce an inverted U-shaped function (the function is monotonically increasing as response probability decreases). Unlike the data, the predicted reaction-time functions (Figure 29) show only slow errors, never errors that are faster than correct responses. Also unlike the data, error responses to extreme stimuli (very high numbers of asterisks or very low ones) are never produced by the model. This is a result of the model's deterministic processing (with variability coming only from the random sequence of stimuli) and the fact that extreme stimuli are almost always consistently assigned to one response.

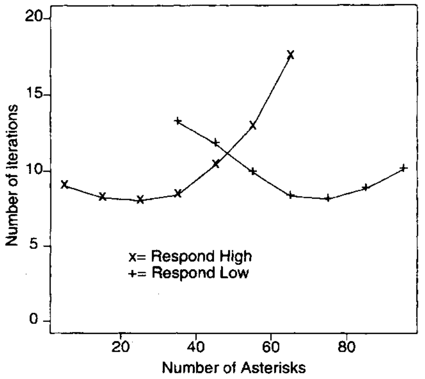


Figure 28. Reaction-time functions predicted from the brain-state-in-a-box model for the data from Experiment 1 (see Figure 2).

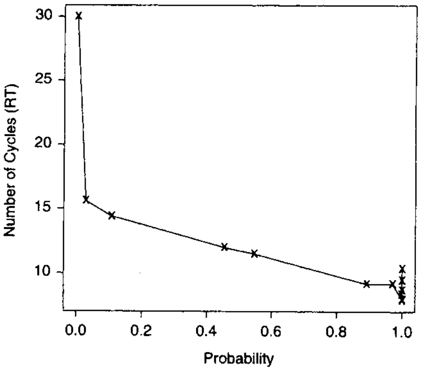


Figure 29. Latency–response probability functions predicted from the brain-state-in-a-box model for the data from Experiment 1 (see Figure 3).

Reaction-Time Distributions

The shapes of the reaction-time distributions produced by the model, shown in Figure 30, are similar to the experimental data, with skewing to the right. Also like the data, the hazard functions (Figure 31) are peaked and fall slowly from the peak or asymptote at the peak.

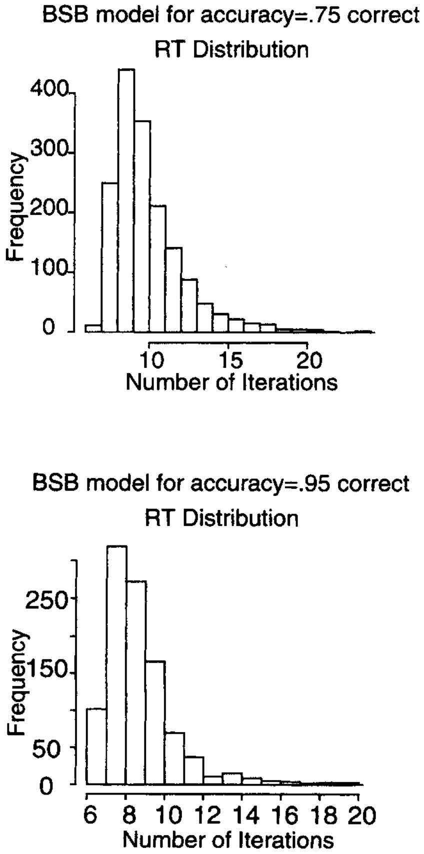


Figure 30. Reaction-time (RT) distributions predicted from the brain-state-in-a-box (BSB) model for the data from Experiment 1 (see Figure 4).

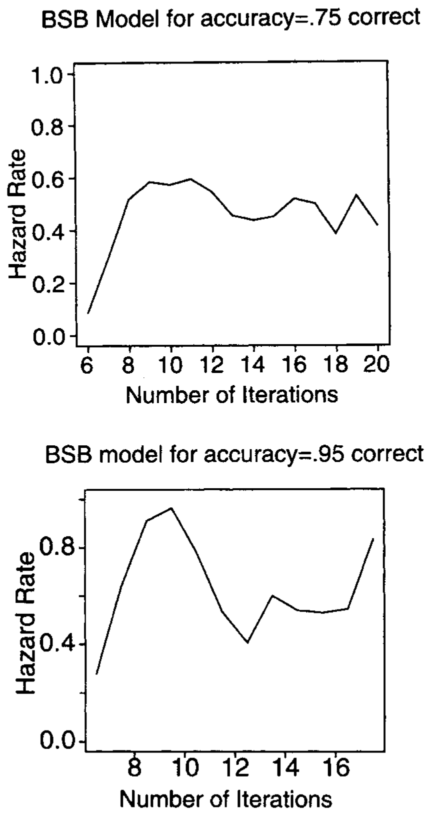


Figure 31. Reaction-time hazard functions predicted from the brain-state-in-a-box (BSB) model for the data from Experiment 1 (see Figure 6).

Learning

Initial learning took about 100 trials before performance reached asymptote in reaction time. Response probability is still quite variable up to 200 trials, but after that performance seems to asymptote.

Adding Variability

One way the BSB model fails is in fitting the inverted U-shaped latency–probability functions, because error reaction times are always slower than correct reaction times. We decided to try adding variability into processing in order to produce a small proportion of fast errors to better mimic the data. To add variability, features in the input representation were allowed to randomly reverse from 0 to 1 or 1 to 0 at the beginning of a cycle of processing. So, for example, if the stimulus was 90 asterisks, reversals could give a 1 in element 1 or a 0 in element 90. The problem with this modification was that feature reversals in extreme elements began to dominate what was learned in the infrequent encounters with these stimuli, and so there was a huge increase in the probability of an error for these stimuli (e.g., up to 50% errors).

Another way we tried to get the model to produce fast but infrequent errors to extreme stimuli was to reverse the response assignment with some small probability. So a stimulus generated from the low distribution was sometimes (10% of the time) associated with high feedback. This did produce a few errors in the extreme tails, but only very slow ones.

We found no reasonable way to get the BSB model to produce a few extreme errors with fast reaction times. Because the model sums products of elements, introducing errors randomly reduces accuracy for extreme stimuli, leading to too many errors on these stimuli.

Summary

The BSB model qualitatively mispredicted error reaction times, and there were problems with sequential effects for some subjects. We were unable to discover any way to modify the model to deal with these failures. This model is very simple and straightforward, and it may be that additional assumptions might allow it to deal better with the experimental data. But we have been unable to find any to this point. The difference between this model and the GRAIN-based models is that the BSB model is more constrained than the GRAIN-based models. For the BSB model to fit reaction-time data, theoretical development of the model is needed.

Experiment 2

A marked feature of the diffusion model's account of the data from Experiment 1 was the match between drift rates in the model and the probabilities that the stimuli were drawn from the high versus the low distributions. Moreover, because with this match the diffusion model fit the data so well, we used stimulus probabilities to train the second GRAIN-based model, and it also gave a good account of the data. Thus, the link between stimulus probability and performance was crucial. Experiment 2 was designed to further test this link by varying stimulus probabilities.

In Experiment 1, stimuli were chosen equally often from the two distributions, high and low, and this was also true for some conditions of Experiment 2; we refer to this as the 50:50, equal bias, condition. The two distributions are shown in the middle top panel of Figure 32, which shows for every possible number of asterisks the probability with which it was drawn from the high and low distributions. The crossover point, 50 asterisks, is the number of asterisks for which the probability it was drawn from the high distribution is .5, equal to the probability that it was drawn from the low distribution.

In most conditions of Experiment 2, stimuli were chosen from the two distributions unequally; either stimuli were chosen from the low distribution with probability .8 and from the high distribution with probability .2 (the low-bias condition, 80:20), or the reverse, from the low distribution with probability .2 and from the high distribution with probability .8 (the high-bias condition, 20: 80). The left and right top panels of Figure 32 show the probabilities for these two bias conditions and the vertical lines (at the crossover points) show the stimulus that was equally likely to have been drawn from the high and low distributions.

Across the course of each experimental session, all three conditions were used—equal, high, and low bias. The questions were whether subjects' performance would track the varying stimulus probabilities, whether they would do so in a way that could be explained by the diffusion and GRAIN-based models, and whether the explanations would be consistent with the explanations of the data from Experiment 1.

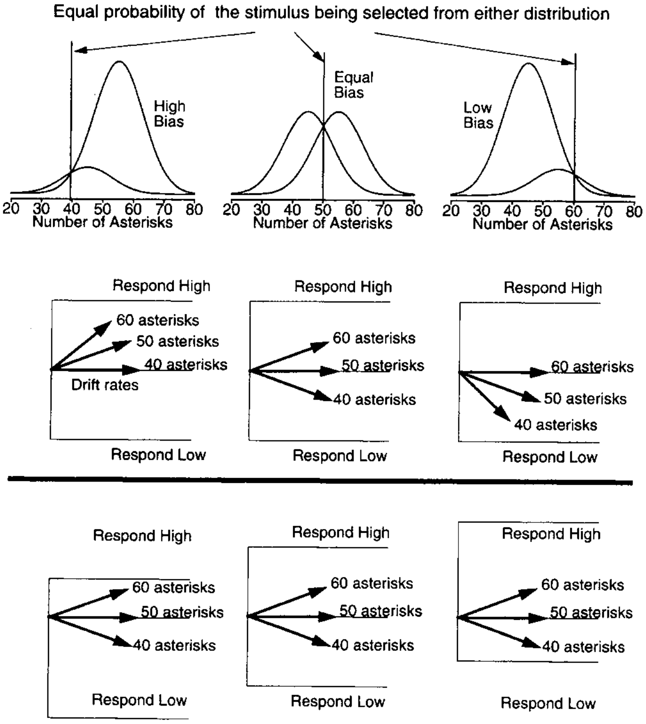


Figure 32. An illustration of two ways the diffusion model can account for changes in stimulus probability. The top row shows three different bias conditions, high, equal, and low, respectively, with a vertical line denoting the equal probability position. The middle row shows the diffusion model with drift rate criterion varying to follow stimulus probability as a function of bias condition. The bottom row shows the diffusion model with boundary positions being altered to account for changes in stimulus probability.

For the diffusion model, a shift in the probabilities of stimuli being chosen from the high versus low distributions could be modeled in three ways: with drift rates changing to accommodate the new probabilities, with boundaries changing, or with both. The model's account of the data of Experiment 1 implicates changing drift rates. What changing drift rates would mean in Experiment 2 is illustrated in the middle row of panels in Figure 32. With the equal bias (middle panel), 50:50, condition, the mean drift rate for 50 asterisks is zero because 50 asterisks was equally likely to be chosen from the high as the low distribution. The mean drift rate for 60 asterisks is positive and the mean drift rate for 40 asterisks is negative. Switching to, for example, the low-bias condition (right panel), stimuli are much more likely to have come from the low distribution so the mean drift rates shift toward low responses: The stimulus with mean drift rate of zero is now 60 asterisks, and both 50 asterisks and 40 asterisks have negative drift rates. Compared with the 50:50 condition, responses to 40 asterisks are, on average, faster and more likely to be low; responses to 50 asterisks are, on average, somewhat faster and more likely to be low; and responses to 60 asterisks are, on average, less likely to be high and slower.

The stimulus value that corresponds to a drift rate of zero can be thought of as a criterion setting between drift rates for high responses and drift rates for low responses. If the number of asterisks in a stimulus is above this criterion value, the mean drift rate will be positive, producing, on average, a high response; if the number of asterisks is below the criterion value, the mean drift rate will be negative, producing, on average, a low response.

The bottom row of panels in Figure 32 shows the second way switches in probability could be modeled for Experiment 2, boundary movement. For the 50:50 condition (middle panel), the boundaries are equidistant from the starting point. For the low-bias condition, for example, the low boundary moves closer to the starting point (and the high boundary could either move away, as illustrated, or remain where it is). As a consequence, low responses for all stimuli become more likely and, on average, faster.

These adjustable criteria have been used to explain the effects of varying instructions and payoffs and other manipulations of response probability. For example, in the 1980s, the diffusion model was the center of a debate about whether the finding that "same" responses were generally faster than "different" responses in same–different matching tasks could be explained by criterion settings or instead required the postulation of an extra stage of processing (Proctor, 1986; Proctor & Rao, 1983; Ratcliff, 1985, 1987; Ratcliff & Hacker, 1981, 1982). Ratcliff (1985) showed that the diffusion model could explain the data without recourse to a separate stage of processing by using adjustments in the zero point of drift and in the response boundary positions.

For Experiment 2, the two possibilities, shift in drift rates and shift in boundary positions, make somewhat different predictions about the shapes of reaction-time distributions. For example, moving a boundary closer to the starting point means that the leading edge of the reaction-time distribution is reduced. The crucial prediction from Experiment 1 is that shifts in probabilities lead to shifts in drift rates, although there is no reason that boundary shifts might not also occur. As it turned out, the diffusion model fit the complete pattern of data best with both drift rates and boundary positions shifting.

Method

Subjects. The subjects were 2 Northwestern University undergraduates (1 man and 1 woman), who were paid for their participation. Both had normal or corrected-to-normal vision.

Stimuli. The number of asterisks to be presented on a trial was drawn from either a low distribution, with mean 45, or a high distribution, with mean 55. The standard deviation of both distributions was 8, giving a d' value of 1.25. The two distributions crossed at the number 50. These distributions were changed slightly from those used in Experiment 1 to give more observations per condition in the central region (e.g., 40–60 asterisks).

Procedure. The procedure was the same as that of Experiment 1, except that a monetary payoff scheme was used to motivate the subjects' performance. Subjects were encouraged to make their responses quickly, although they were told that their goal should be to maximize the total number of points earned over the course of the experiment. Four points were awarded for every correct response, and one point was subtracted for every incorrect response. Subjects were paid at a base rate of $6 per session and told that their pay would be supplemented according to the total number of points that they earned: $.70 for every 1,000 points. Thus, for a block of 40 trials, a subject could earn as many as 160 points or an additional 11 cents toward their pay.

Design. Each subject performed in 10 sessions over approximately 2 weeks. Each session was composed of 30 blocks of 40 trials each. In all sessions, rest breaks were inserted every 2 blocks, as in Experiment 1.

In the first two (practice) sessions, the stimuli were drawn from the low and high distributions with equal probability (50:50, as in Experiment 1). In the remaining eight sessions, the blocks were organized as follows: First, there were four blocks for which the probability of choice from the two distributions was 50:50, as in the practice sessions. Then there were four sets of blocks to implement switches from one bias condition to the other. Each set began with one 50:50 block, and then there were one, two, three, or four blocks all either high bias (20:80) or low bias (80:20); then the opposite bias was used for three blocks. Across the four sets, each possible number of blocks preceding a switch (one, two, three, or four) was used once. In two of the sets, the switch was from high to low bias, and in the other two, it was the reverse. The order of the sets and the assignment of the direction of the switch to each set was random, except for the constraint that, overall, there had to be an equal number of high-to-low and low-to-high switches for each of the possible numbers of blocks preceding switches.

Results

In the data analyses, all trials with response times less than 170 ms or greater than 3000 ms were discarded (less than .1% of the data). High- and low-bias blocks of trials were collapsed across each other because the experimental results were symmetrical. This was done by subtracting the number of asterisks presented on a high-bias trial from 100 and reversing the subject's response. This allowed high responses from the high-bias condition and low responses from the low-bias condition to be combined. This meant that data could be presented in terms of "preferred" responses (high responses in the high-bias condition and low responses in the low-bias condition) and nonpreferred responses (low responses in the high-bias condition and high responses in the low-bias condition). For example, for the stimulus 60 in the high-bias condition, the preferred response was "high," and for the stimulus 40 in the low-bias condition, the preferred response was "low." Data for the stimulus 60 in the high-bias condition with the response "high" were combined with data for the stimulus 40 in the low-bias condition with the response "low" to produce an average response probability and response time. Similarly, data from the stimulus 60 in the high-bias condition with the response "low" were combined with data for the stimulus 40 in the low-bias condition with the response "high" to produce an average response probability and response time.

Asymptotic Performance and the Diffusion Model

The data from Experiment 2 present two important tests of the diffusion model: Can it account for the asymptotic response latency–probability functions and reaction-time distributions as well for Experiment 2 as it did for Experiment 1, and can it provide the same probability matching explanation for drift rate for individual subjects as in Experiment 1?

The data from the second and third blocks of trials after a switch from one probability condition to another were used in the analyses for asymptotic performance. The diffusion model was fit to the data in the same way as for Experiment 1, choosing three stimulus conditions to fix the model parameters and then generating predictions from those parameters for the full range of conditions. There were about one half the number of observations per subject as in Experiment 1, and so there is more variability in the data. The latency–probability functions, both data and model predictions, are shown in Figure 33. The top figures show the latency–probability functions for the nonpreferred response, and the bottom figures show the latency–probability functions for the preferred response. In the top panels, responses to the far right correspond to errors, as they were defined in Experiment 1. For both subjects, there is a sizable difference in reaction time between preferred and nonpreferred responses (100–200 ms), but the reaction-time functions are relatively flat as a function of response probability. The model accounts for these trends reasonably well, capturing the error versus correct reaction times, even though the differences are small. The model also does a good job with the reaction-time distributions (Figure 34), except for preferred responses for Subject 2. The parameters of the model are shown in Table 1.

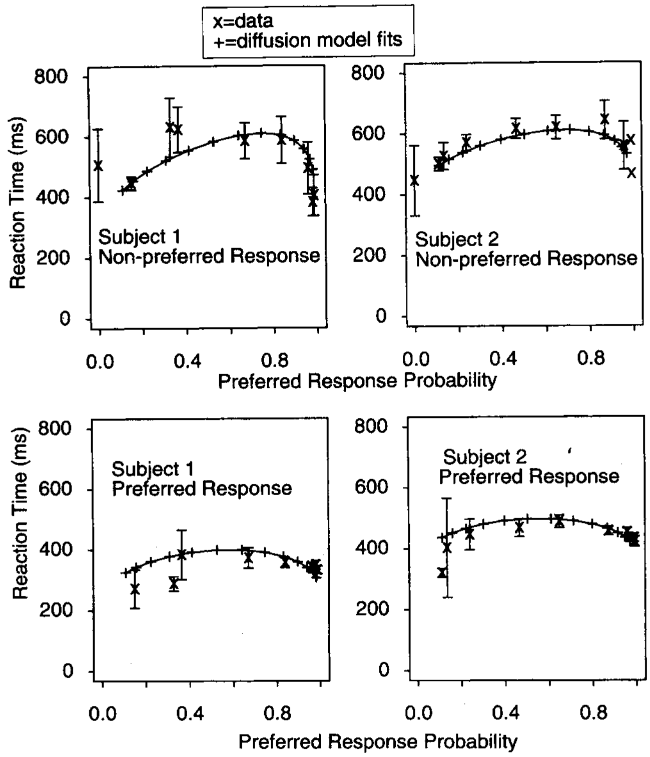


Figure 33. Latency–probability for the two subjects in Experiment 2 for high- and low-probability responses and predictions from the diffusion model. The error bars represent 2 standard deviations.

Given that the model fits the data reasonably well, we can ask what the model has to say about how subjects adjusted their asymptotic behavior to deal with unequal stimulus probabilities. One way they adjusted, as shown by the parameters in Table 1, was to move the starting point closer to the boundary for the preferred response than to the boundary for the nonpreferred response (cf. Ratcliff, 1985).

The second way they adjusted was to shift drift rates to conform to the new probabilities with which stimuli were chosen from the high versus low distributions. Just as in Experiment 1, subjects' drift rates matched the probabilities that stimuli were chosen from the high versus the low distributions. The functions in Figure 35 show this matching. The drift rates were plotted as a function of number of asterisks, with data from the low-bias condition combined with the data from the high-bias condition by flipping the

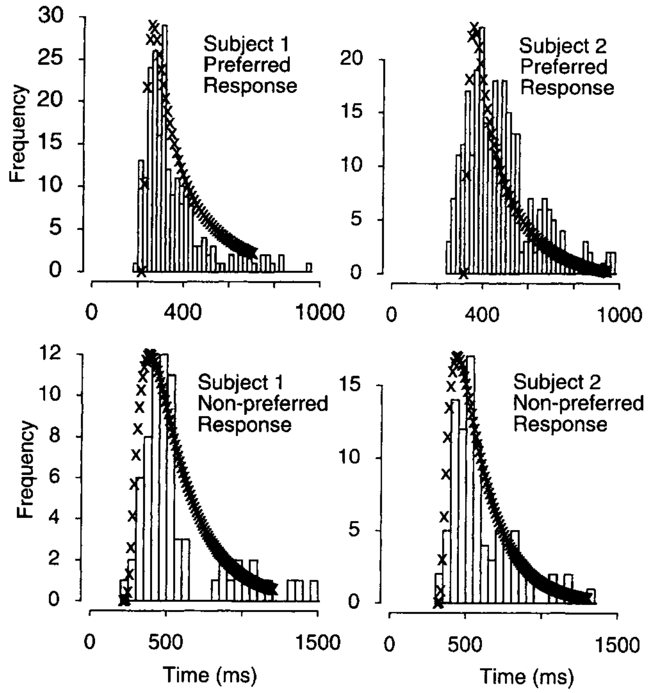


Figure 34. Sample reaction-time distributions for the two subjects in Experiment 2. Predictions from the diffusion model are shown by the asterisks.

number of asterisks scale around the midpoint 50 (so 45 and 55 would be combined, as above). Both subjects' drift rates lie virtually on top of the probability function for the high-bias condition (with probability transformed as in Experiment 1). The theoretical probability functions for the high-bias condition (also representing the flipped low-bias condition) and the 50:50 condition were derived from the density functions that controlled the assignment of feedback to responses; these functions are shown in the bottom panel of Figure 35 (cf. Figure 32). For example, in the 50:50 case, the crossover point of the two distributions was 50, whereas for the high-bias condition, the crossover point was about 60. This shift is reflected in the probability curves in the top panel where the midpoints of probability (the zero point on the left-hand axis) correspond to 50 and 60 asterisks, respectively.

Sequential Effects in the Diffusion Model

In Experiment 1, subjects showed different patterns of sequential effects: Two showed a greater probability of responding in the direction of the prior response, one showed the opposite effect, and one showed no sequential effects. We postponed discussion of these sequential effects to this point because it turned out that they can be explained with the same parameter changes that account for the effects of changes in the probabilities with which stimuli were drawn from the high versus low distributions.

The sequential effects shown in Figure 1 cannot be produced with changes in the starting point z alone. Changes in the starting point of ±15%, 10%, 2%, and 10% of z, for Subjects 1 through 4, respectively, would give the correct differences in response probabilities (.25, .12, .03, and .17, respectively). But these 10% or 15% changes give 100–200-ms changes in response time, far too large in comparison to the data. For example, for reaction time, for the middle two points on the response probability function in Figure 1, the differences in reaction times were 44 ms, −51 ms, 20 ms, and 10 ms faster for repetition of the same response, as opposed to switching from the opposite response, for Subjects 1 through 4, respectively.

Shifts in both starting point (equivalent to a shift in one boundary away from the starting point and one boundary toward the starting point) and drift rate are needed. A relatively small change in drift rate (.05 for Subject 1, e.g.) produces changes in response probability of the magnitude shown in Figure 1 (see also Ratcliff, 1985, 1987), but alone it produces little change in reaction time. When a small shift in drift rate is combined with a small change in starting point, the model accurately fits sequential effects in both accuracy and response time. Thus, sequential effects are explained by the same mechanisms that account for subjects' responses to a change in the probabilities with which stimuli are drawn from the high versus low distribution.

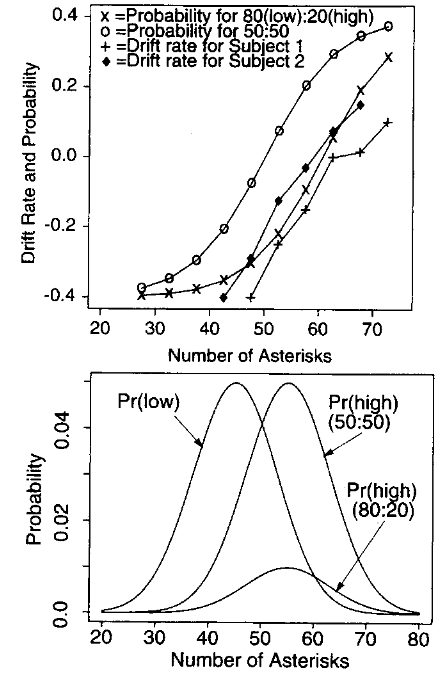


Figure 35. Top: Drift rate estimated from the data for the two subjects in Experiment 2. Data and fits are for the average of high responses to high stimuli when the stimulus is selected from the high distribution and the high distribution is more probable, and low responses to low stimuli when the stimulus is selected from the low distribution and the low distribution is more probable. Also shown are the probability of a high response for the 50:50 bias and the 80:20 bias conditions. These are transformed from the range 0 to 1 to the range −.4 to +.4 by subtracting .5 from probability and multiplying by .8. The drift rates follow probability that a stimulus is selected from the most probable distribution. Bottom: Density functions for the stimuli in Experiment 2 where the two tall curves represent the 50:50 condition and where the small curve represents the low-bias condition in which stimuli are selected from the high distribution 20% of the time. The two probability curves in the top panel are derived from the bottom curves. Pr = probability.

In general, the diffusion model uses shifts in drift rate to accommodate sequential effects in response probability that are accompanied by small or no changes in reaction time as shifts in drift rate. The model uses shifts in boundary positions to produce large sequential effects in reaction time and smaller effects in response probability. In the data presented here, shifts in both boundary position and drift rate are needed to greater or lesser degrees for the four different subjects in Experiment 1.

Sequential Effects in the GRAIN-Based Model

In the GRAIN-based model, with all learning prior to the experiment, there is no learning mechanism during the experiment to model sequential effects or changes in stimulus probability, so they must be modeled with assumptions similar to those used in the diffusion model. As with the diffusion model, a combination of altering the response criteria and altering the way a stimulus is interpreted is sufficient. We illustrate this for the sequential effects. We examined three possibilities: First, instead of resetting activation in the output nodes to zero after each trial, some portion of activation in the output nodes could be carried through to the next trial, and, in the case of Subject 2 in Experiment 1, the sign of the activation would be reversed (because the sequential effects are reversed relative to the other subjects). We implemented this idea, setting the amount of activation to be carried forward to about two thirds of the final value on a trial, producing appropriate changes in response probability but changes that were much too large in response time. For example, assuming a scaling factor of 5 ms of response time per activation cycle, we found a 150-ms reaction-time effect for a .06 change in probability. In contrast, the effect for Subject 1 was 44 ms in response time and .25 in response probability.

A second possibility is that the response criteria vary as a function of the prior response in the same way that boundary positions varied in the diffusion model. With a 4% change in response criteria, the reaction-time difference was about right, 40 ms, but the accuracy difference was too small (.05).

The third possibility is to assume that when a response is produced, there is a bias to interpret the next stimulus the same way as the previous stimulus (for Subjects 1, 3, and 4, or the opposite way for Subject 2 in Experiment 1). This can be done with an input level node that takes as its input the output from the prior trial. The effect would be the same as shifting the drift rate function in the diffusion model (see Experiment 2). A combination of these last two assumptions (response criteria and drift rate bias) produces the correct relative sizes of the changes in reaction time and accuracy to fit the data, and the two assumptions mimic the assumptions used in fitting the diffusion model. This combination also accounts for the asymptotic effects when stimulus probability is manipulated as in Experiment 2.

Adaptation

The discussion of the results of Experiment 2 so far has been limited to asymptotic performance. In this section, we look at adaptation from one bias condition to another. Figure 36 shows that subjects were able to adapt very rapidly; specifically, it shows the probability with which subjects gave the preferred response after a switch between high- and low-bias conditions. The three curves divide the stimuli into groups of 1 to 39 asterisks, 40 to 60 asterisks, and 61 to 99 asterisks. For the middle range (40 to 60), subjects should, most of the time, give the response toward which trials were biased (high in the high-bias condition, and low in the low-bias condition). This is what they did; the asymptotic probability of the preferred response was .8. Figure 36 shows that they reached this asymptotic probability within about 14 trials after a switch. Before the switch, the probability of what was to become the preferred response was .2 (1–.8), and even after only 5 trials after the switch, the probability of the preferred response had moved two thirds of the way toward asymptote (i.e., to about .6).

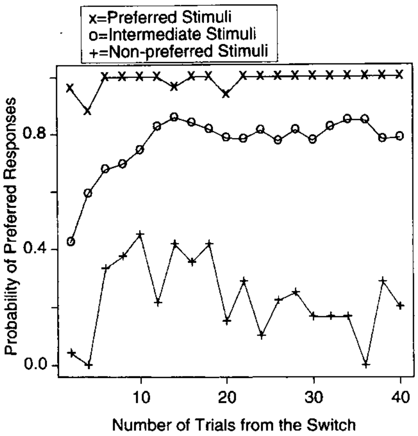


Figure 36. Response probability as a function of the number of trials from a switch from one bias condition to the other (e.g., 80:20 to 20:80). For the high-bias conditions, preferred stimuli (Curve 1) are 61–99 asterisks, intermediate stimuli (Curve 2) are 40–60 asterisks, and nonpreferred stimuli (Curve 3) are 1–40 asterisks. For the low-bias condition, preferred stimuli (Curve 1) are 1–40 asterisks, intermediate stimuli (Curve 2) are 40–60 asterisks, and nonpreferred stimuli (Curve 3) are 61–99 asterisks. The asymptotic values of the three curves before the switch were for Curve 1, about .8; curve 2, about .2; and Curve 3, about 0 (1.0 minus the asymptotic probabilities in this figure).

The stimuli that should always be given the preferred response (1–39 in the low-bias condition and 61–99 in the high-bias condition) moved from an asymptotic probability of .2 prior to the switch to reach probability 1 after the switch in four or five trials (see Figure 36). The nonpreferred stimuli moved from their asymptote before the switch of 0 to their asymptote after the switch, .2, in about five trials. Thus, adaptation was rapid, with only five trials required to move most of the way to asymptote.

In sum, the two subjects in this experiment were able to switch from one bias condition to another within only a few trials. To check that this fast learning was not unique to these subjects, we scheduled 32 subjects in a single short experimental session to look at how fast they learned initial probability assignments, using exactly the same experimental design, procedures, and 50:50 probability condition as in Experiment 1. There were five blocks of 32 trials each. Averaging over the 32 subjects, the response probability function for the first 20 trials had almost the same shape as that for the last block of trials (and the function was similar to that for subjects in Experiment 1). The latency–probability functions had roughly similar shapes over the five blocks of trials. The only differences were that responses speeded up a little from the first block to the last block, and the first block did not show any fast errors to extreme stimuli. In short, subjects learned to probability match in the first 20 trials.

Neither the diffusion model nor the GRAIN-based model with learning prior to the experiment has any mechanism for learning during the experiment, and so, of course, neither can give an account of the speed with which subjects switch from one bias condition to another. To examine whether the BSB model or the GRAIN-based model with learning during the experiment could adapt to switches as rapidly as subjects did, we first trained the models to stable performance using the same parameters as for Experiment 1. (This means that the models have the same problems in accounting for latency–probability functions as they did in Experiment 1, but the issue here is their ability to account for adaptation.)

Once the models were trained to mimic a low-bias condition, we chose a stimulus that would have received mainly high feedback in the 50:50 condition (i.e., 65) and presented it to the model with low feedback for several trials. We trained the model with this stimulus with low feedback until the model produced mainly low responses to stimuli below 80 and mainly high responses to stimuli above 80 (i.e., a crossover point of about 80). Then to mimic switching to the high-bias condition, we presented the stimulus 40 for several trials, each trial with high feedback, allowing the model a chance to learn to respond mainly high to stimuli above 40. After each trial of stimulus 40, we checked all 96 stimuli 10 times each to see what average response the model gave to each.

For the BSB model, after one learning trial with the stimulus 40, the crossover point had moved from 80 to 72; after the next five trials, it moved to 64, 56, 36, 24, and 16, respectively (note that the sequence does not asymptote because the sequence of trials was not randomly chosen). Thus, in four trials, the BSB model had moved about two thirds of the way toward a crossover point of 20 (symmetrical with 80), approximating the course of adaptation for the subjects (as shown in Figure 36). The BSB model predicts performance well because it essentially computes a running average of input plus feedback, and the weighting it places on current input relative to prior input determines how quickly it adapts to changes in probability. The parameters of the model that account for asymptotic behavior also account for the rapid changes in performance that come about when the probability of high versus low feedback for stimuli is changed.

The GRAIN-based model, with the parameter values used to produce the fits in Experiment 1, produced more rapid adaptation. On the second trial, the crossover point had moved two thirds of the way toward 20, adapting much faster than the human subjects. These changes are too large to mimic the adaptation found in the data (cf. catastrophic interference; McCloskey & Cohen, 1989; Ratcliff, 1990).

The GRAIN-based model might possibly do better by adding some kind of bias to processing, for example with an input node set to represent bias (e.g., Cohen et al., 1990). The node would be weighted according to bias conditions and so weight one response over another. But this scheme would require external intervention to tell the system when bias had changed.

Although the information we provide here about rate of adaptation is limited to one experiment with two subjects, it does point the way to a set of issues that connectionist models must address, suggesting that further theoretical development should seek common mechanisms for adaptation and sequential effects.

Summary

The diffusion model accounted for the data of Experiment 2 by modeling changes in the probabilities with which stimuli are chosen from the high versus low distributions as changes in drift rates and boundary positions. The changes in drift rate follow stimulus probability as predicted from the results of the fits of the model to the data from Experiment 1. Altering both drift rates and boundary positions is consistent with earlier applications of the model to data from letter-matching experiments (Ratcliff, 1985). The model accounts for sequential effects with adjustments to the same parameters, drift rates, and boundary settings as for changes in stimulus probability. The GRAIN-based model with learning prior to the experiment also required changes in those of its parameters that mirror drift rate and boundary positions in order to account for stimulus probability changes and sequential effects. The parallel between stimulus probability effects and sequential effects is reasonable because the adaptation to the changes in stimulus probability was rapid, occurring within about 5 to 10 trials, consistent with immediate, one trial to the next, sequential effects.

General Discussion

The aims of the research presented in this article were to evaluate how well the connectionist models and the diffusion model could account for reaction-time phenomena. The success of the enterprise is demonstrated by a number of significant new findings, enumerated in the paragraphs below. Some of the new findings are pertinent to the individual models. The connectionist models, for example, were put in contact for the first time with all the measures that have been traditionally used in reaction-time research. For the diffusion model, new insights were gained by the model's application to a new paradigm. Others of the new findings are more general. For one, we had not anticipated using insights gained from the diffusion model to choose a training function for connectionist models. For another, we had not appreciated how severe were the constraints imposed on the connectionist models by joint consideration of reaction time and accuracy. Perhaps the most significant outcome is the platform provided for future research. To explain reaction-time phenomena up to the standard set by the diffusion model, new models must explain and explicitly fit correct and error reaction times, the shapes of reaction-time distributions, and accuracy. The adaptation results from Experiment 2 are preliminary but point to the need to place all the models in a more general framework that can explain learning and adaptation phenomena and the mechanisms by which decision criteria are set. Overall, a comprehensive and difficult research program is laid out for both connectionist and more traditional reaction-time models.

The Diffusion Model

The success of the diffusion model in explaining the empirical data was a pleasant surprise. With only five or six parameters for each subject, the model accurately fit correct and error reaction times, their probabilities, the shapes of their distributions, and their hazard functions. Previously, no model had been able to explain the relative speeds of correct and error responses, so the diffusion model's ability to do this is a significant advance.

The diffusion model offered insights into the behavior of the individual subjects. Most important, the model showed how their behavior could be related to stimulus probability. Although intuition might suggest that subjects should match their responses to the probabilities of the stimuli coming from the high versus low distributions, there was nothing in the data themselves that directly showed subjects did so. Only by the model's extraction of drift rates from the response-time and response-probability data was it discovered that the underlying variable was stimulus probability. We also found that subjects' shifts from trial to trial within a probability condition and shifts from one probability condition to another could be explained by the same changes in behavior in the model: changes in boundary positions and changes in the zero point of drift rate (drift criterion). Thus, fitting the diffusion model led to explanations of the bases of subjects' decision making.

The model also provided an understanding of how the individual subjects could rely on the same underlying information as the basis for their decisions, yet produce quite different speed–accuracy profiles (see also Ratcliff & Rouder, 1998, Experiment 1). In our Experiment 1, responses from two of the subjects were slow overall, responses for another subject were fast overall, and responses for the fourth subject were intermediate in speed. According to the model, these differences came about from differences in how far response boundaries were set from the starting point. The subjects also differed in the speed of correct versus error responses: One subject showed slightly faster errors than correct responses, one subject showed error responses slower than correct responses, and two subjects showed errors to extreme stimuli faster than correct responses and errors to less extreme stimuli slower than correct responses. According to the model, these different patterns are due to different amounts of variability in how the stimuli were encoded from one trial to the next and different amounts of variability in the starting point of the decision process from trial to trial. The subject with slightly faster error than correct response times had little variability in encoding across trials; the other subjects had more. Three of the subjects kept the distances of the boundaries from the starting point about constant across sessions, whereas one subject reduced them.

The success of the diffusion model provides an impetus for adding variability to parameter values in other models. For example, any model that predicts an inverted U-shaped latency–probability function (i.e., errors for extreme stimuli faster than errors for nonextreme stimuli) could also predict an asymmetric function if variability was added to the appropriate parameters (see Audley & Pike, 1965). Then, just as in the diffusion model, error reaction times would become a mixture of slower processes with higher probability and faster processes with lower probability, and the average of the mixture would be slower than for a single process with the same drift rate as the mean of the mixture.

In sum, the diffusion model provided a complete account of all of the measures of the time course of processing involved in decisions about whether the number of asterisks in a stimulus was high or low. We are optimistic that the model can provide an equally good account for other rapid, cognitive, binary choice decisions (e.g., Ratcliff & Rouder, 1998, showed generalization to perceptual discriminations, and Ratcliff & Rouder, in press, showed generalization to letter identification with masking) and so provide a theoretical basis across a range of specific experimental paradigms.

Signal Detection Theory and the Diffusion Model

The most popular description of response probability in two-choice tasks is signal detection theory. It is important to understand how the signal detection description differs from that offered by the diffusion model; here we present two examples of differences.

The diffusion model offers an explanation of behavioral effects on two dependent variables, accuracy and response time, whereas signal detection theory deals only with accuracy. In this sense, the diffusion model can be thought of as an extension of signal detection theory to the domain of reaction time (e.g., Ratcliff, 1978). However, there is an important but often unappreciated problem with signal detection theory, namely that it does not produce invariant measures of d's across experimental conditions for which d' should be constant. The diffusion model, on the other hand, can produce constant d's. For example, when subjects shift their speed–accuracy criteria in one condition of an experiment versus another, signal detection theory will produce different d' values for the two conditions. This is true even when nothing about the stimulus has changed between conditions and nothing about the task has changed except for whatever manipulation (e.g., instructions) causes subjects to be faster in one condition and more accurate in the other. But in the diffusion model, differences in behavior between the conditions can be modeled by differences in boundary positions, leaving d' invariant (see Ratcliff & Rouder, 1998, Experiment 2).

The situation becomes more complicated when the probability of making one response relative to another varies across conditions (as it did across the high- and low-bias conditions of Experiment 2). Signal detection theory explains the differences in behavior across the conditions as a change in criterion; d' might or might not be constant across conditions depending on how subjects adapt to the different biases. In the diffusion model, the changes in relative probability of the responses might be produced by changes in the drift criterion, by changes in the starting point of the diffusion process, or by one boundary moving nearer the starting point. In each of these cases, the diffusion model would produce the same value of d' across the conditions. The differences in the three ways of producing changes in response probability would show up in the shapes of the reaction-time distributions. Generally, the diffusion model and other sequential sampling models are good candidates to serve as generalizations of signal detection theory to the reaction-time plus accuracy domain.

Mathematical and Biological Plausibility of the Diffusion Model

It is well-known in the stochastic process literature that, in any domain, the diffusion process has defects as a model of processing because the velocity in the process becomes infinite when time steps become very small (Cox & Miller, 1965, p. 207). The Ornstein–Uhlenbeck process (Cox & Miller, 1965; Smith, 1995) has sometimes been used as an alternative because it does not have this problem. However, another way to avoid the problem is to think of the diffusion process as a continuous approximation of discrete neural events that take place in the range of milliseconds. A discrete process with steps of about 1 ms does not have the problem of infinite velocity, and it is well approximated (to within a percentage or two) by the continuous process. Thus, with the understanding that the diffusion process is an approximation to neural events that are on a time scale of milliseconds, it is computationally and biologically plausible. In fact, the diffusion process has a long history of service as a model of the stochastic behavior of single neurons (much of the monograph by Tuckwell, 1989, is devoted to discussion of diffusion processes). Moreover, recent work by Hanes and Schall (1996) has shown that the firing rates of cells in the frontal eye fields of monkeys (cells responsible for moving the eyes) appear to accumulate information up to a fixed response criterion just as is described by sequential sampling models. This link between the models and neural behavior offers a possible avenue for future convergence of behavioral models and neural functioning.

Connectionist Models

Only one of the three connectionist models that we examined could explain all the response-time and accuracy phenomena. The BSB model could not produce the right error response times (nor could it learn the task as quickly as could subjects). The first of these problems might be fixed by adding variability to the model (although we could not find a way to do this), but speeding up learning to address the second problem might mean changing the model altogether to somehow incorporate prior knowledge. The GRAIN-based model with learning during the experiment had many problems: It could not accurately produce sequential effects, the relations between correct and error response times, or adaptation from one stimulus probability condition to another. We could find no variations in parameter values or architectural assumptions to remedy these problems. The GRAIN-based model with learning prior to the experiment could fit response-time and response-probability data (though some hazard function fits missed significantly), but it could not adapt from one stimulus probability condition to another without additional assumptions. In this respect, it is similar to the diffusion model. In fact, the GRAIN-based model with learning prior to the experiment and the diffusion model can be regarded as approximations of each other; this came about because we used the diffusion model as a guide for training the GRAIN-based model.

Our explorations of the connectionist models found none of them completely successful. Nevertheless, we believe there are several important findings. One is that the connectionist models we considered are aiming at a very high hurdle. They are intended to describe how decisions are reached over time and in doing so account for all the empirical measures of decision processes, plus learning and adaptation. Our findings suggest that it will be very difficult for a model to move from an account only of mean response time or only of mean probability of one response versus another to a much fuller account.

A second important finding arising out of our attempts to have connectionist models explain response-time phenomena was a better understanding of a problem we had not originally appreciated, namely the interaction between the two kinds of measures, learning and response time. Many connectionist models learn from feedback; on each trial, the model makes a response and then updates the weights in the network on the basis of feedback about whether the response was correct. Thus, the decision processes are constrained at learning only by feedback. Holding the model responsible not only for response probabilities, but also for response times, a second dependent variable, and all the attendant response-time measures (e.g., the shapes of response-time distributions and error-response times) imposes constraints to which the learning algorithm has no access. This means that feedback alone (which may be sufficient to allow the model to fit response probabilities) will not be sufficient to produce fits to reaction-time measures. Adding to the complexity of this problem is that connectionist models often assume a distributed representation, which means that the model has to be trained to respond to the whole stimulus set at once. Putting this all together, for a connectionist model with learning built into processing during an experiment, deriving fits of the model to data involves, first, training the model on a random sequence of the stimuli and, as trials proceed, producing response-time and accuracy predictions from the responses the model makes to the stimuli; then second, comparing the predictions and empirical data to adjust the parameters of the model; and then repeating this cycle until the predictions and data match. For a model with training prior to the experiment's sequence of trials, the weights on connections among nodes would be set by training the model to produce some functional mapping between stimuli and output. Once the training phase was complete, testing would be carried out across simulated trials to produce predictions for response measures, these would be compared with the empirical data, the parameters of the model would be adjusted, and the training–testing cycle would be repeated until the best fit was obtained. With current workstations, the whole process of finding the best possible fit of a connectionist model to data would take weeks for data such as those from the asterisk task. It is for this reason that we fit the connectionist models to data by hand, not with automatic search algorithms.

Another issue we had not initially appreciated was the difficulty and complexity of the problems imposed by requiring the models to learn and to adapt to changes in stimulus conditions. Subjects learned the asterisk task much too quickly for the models; hence, the models cannot start with a completely blank slate, learning only from feedback on the first few trials. But how could preexisting knowledge be built in? One possibility is to combine or select from preexisting networks, but how to choose and calibrate such networks is an issue that has not been tackled in connectionist research. Another possibility might be to incorporate some knowledge into the structure of a model and then let the model learn whatever extra knowledge was required for a particular combination of stimuli and task. Again, this approach has not yet received serious examination in this domain.

Of course, we have explored only a tiny portion of the space of connectionist models. GRAIN provides a set of principles for model construction from which we constructed only two models (and some of their variants). Many other learning rules, representations, or assumptions about the cycling of activation could be tried; the number of possible models is very large. Beyond GRAIN are many other kinds of neural–connectionist models, most of them not yet addressing reaction-time phenomena. In this article, we have barely scratched the surface of application of neural–connectionist models to reaction-time phenomena.

Comparing the Diffusion Model and Connectionist Models

We view the diffusion model as a general-purpose decision mechanism; it can be used in a variety of situations and tasks, and the decisions it models can be based on any one of a variety of kinds of information. For example, random walk and diffusion models have been used in simple and choice reaction time (Laming, 1968; Link, 1975; Link & Heath, 1975; Smith, 1995; Stone, 1960), letter matching (Ratcliff, 1981, 1985), discrimination and same–different tasks (Ratcliff & Rouder, 1998), recognition memory (Ratcliff, 1978, 1980, 1988), categorization (Nosofsky & Palmeri, 1997), word identification in implicit memory (Ratcliff & McKoon, 1997), and decision making (Busemeyer & Townsend, 1993).

The diffusion process provides a general mechanism that, through application to experimental data, produces a value of drift rate for each experimental condition. The drift rates then provide a measure of the latent variable driving the decision process (so long as the diffusion model accurately fits all the response-time–accuracy aspects of the data). In this way, competing theories about the latent variable can be tested. For example, suppose according to some theory, in some experiment, subjects were basing their decisions on a single dimension of the stimuli, distance from a criterion. The diffusion model would extract drift rates from the accuracy and reaction-time data, and drift rates would provide the functional form of the mapping between distance from the criterion and drift rate. But if the theory was wrong and subjects were basing their decisions on, for example, stimulus probability, the drift rates extracted would correspond to stimulus probability, not distance from criterion. This would falsify the theory (except, of course, in situations in which stimulus probability and distance from the criterion exactly mimic each other).

The diffusion model allows the precise shape of the drift rate function to be extracted from the empirical data. Even if the decision process is not stationary (i.e., it is not constant over the time course of processing; e.g., Ratcliff, 1980; Smith, 1995), it is still possible to examine the drift rate at different points along the time course using deadline or response signal methods and to use drift rate in developing models of dynamic changes in the information driving the decision process.

The important practical point for contrast with connectionist models is that what the diffusion model allows is working backwards from the response time and accuracy data for each individual condition to a characterization of the values of the stimulus dimension that is driving the decision process. This is what the connectionist models we examined did not do.

In the connectionist models, there is no isolable part of the model that corresponds to an individual experimental condition. The reason is that all of the stimuli from all of the conditions must be learned at once, which means that all the information about them is represented collectively (in the weights in the network), not individually. Also, a motivating principle of many connectionist models is interactivity of the various levels of processing (e.g., features, letters, and words). The different levels are usually not designed to be broken apart into processing stages so that the decision stage cannot be separated from the computation of activation that arises from input of the stimulus. The consequence of these two points is that there is no way to work backwards from the individual response-time and accuracy data to derive something equivalent to drift rate to represent the latent variable driving the decision process. Instead, the model has to be fit to data in a forward direction: First, a dimension on which to represent the stimuli must be chosen as well as some way to represent each stimulus on the dimension; second, the model must be trained to the task; and third, stimuli must be processed through the model to give predictions to evaluate against empirical data. In the asterisk task, for example, suppose we did not know whether subjects based their decisions on distance from the criterion or stimulus probability. Then the representation of the stimuli at input would have to be flexible enough to accommodate both possibilities. If the form were too restrictive, then the model might not fit. Then it is hard to see how the failure would point to the correct functional form of the training function.

Conclusions

The diffusion model for two-choice decisions accounts for a wide range of experimental data including response probabilities, correct and error response times, and the shapes of response-time distributions. The model sets a standard against which competing theoretical schemes can be tested. The diffusion model is designed as a general-purpose decision mechanism that can be applied across a wide variety of tasks that require a single rapid binary decision. It is a mechanism that can be pointed at different sources of evidence as a function of task requirements. How it does this and how decision criteria are set are topics for further research.

We hope our evaluation of connectionist models will serve as a challenge to spur development of models to fit the full range of experimental data. However, we anticipate that such development will be difficult because our efforts have shown the complexity involved in efforts to move from a simple account by which mean reaction time is mapped onto some single output quantity of a model to a full account of response time and accuracy data, simultaneously explaining learning.

Evaluating connectionist models is not an easy task because the models are simulations of parallel processes involving nonlinear transformations. The problem is compounded by the large amount of computer time (hours or even days) required for simulations to produce optimal fits. Despite these difficulties, we found it possible to lay out a plausible range of assumptions about representation and process and, in so doing, reveal problems with the constellations of assumptions in the models we tried. An obvious criticism is "you should have tried some other possible models such as . . . ." We hope that others are encouraged to explore these possibilities, using our data and conclusions as a starting point, with the aim of constructing successful connectionist models.

References

Anderson, J. A. (1973). A theory for the recognition of items from short memorized lists. *Psychological Review, 80,* 417–438.

Anderson, J. A. (1991). Why, having so many neurons, do we have so few thoughts? In W. E. Hockley & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock* (pp. 477–507). Hillsdale, NJ: Erlbaum.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning:

Some applications of a neural model. *Psychological Review, 84,* 413–451.

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

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

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

Atkinson, R. C., Bower, G. H., & Crothers, E. J. (1966). *An introduction to mathematical learning theory.* New York: Wiley.

Audley, R. J., & Pike, A. R. (1965). Some alternative stochastic models of choice. *The British Journal of Mathematical and Statistical Psychology, 18,* 207–225.

Bertelson, P. (1961). Sequential redundancy and speed in a serial two-choice responding task. *Quarterly Journal of Psychology, 13,* 90–102.

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 and Psychophysics, 32,* 117–133.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic–cognitive approach to decision making in an uncertain environment. *Psychological Review, 100,* 432–459.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review, 97,* 332–361.

Cox, D. R., & Miller, H. D. (1965). *The theory of stochastic processes.* London: Methuen.

Dosher, B. A. (1984). Discriminating preexperimental (semantic) from learned (episodic) associations: A speed–accuracy study. *Cognitive Psychology, 16,* 519–555.

Espinoza-Varas, B., & Watson, C. (1994). Effects of decision criterion on latencies of binary decisions. *Perception and Psychophysics, 55,* 190–203.

Estes, W. K. (1957). Of models and men. *American Psychologist, 12,* 609–617.

Estes, W. K. (1964). Probability learning. In A. W. Melton (Ed.), *Categories of human learning* (pp. 89–128). New York: Academic Press.

Estes, W. K. (1995). Response processes in cognitive models. In R. F. Lorch & E. J. O'Brien (Eds.), *Sources of coherence in text comprehension* (pp. 51–71). Hillsdale, NJ: Erlbaum.

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.

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

Glaser, R. E. (1980). Bathtub and related failure rate characterizations. *Journal of the American Statistical Association, 75,* 667–672.

Gronlund, S. D., & Ratcliff, R. (1989). The time-course of item and associative information: Implications for global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 846–858.

Hacker, M. J. (1980). Speed and accuracy of recency judgments for events in short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 6,* 651–675.

Hanes, D. P., & Schall, J. D. (1996). Neural control of voluntary movement initiation. *Science, 274,* 427–430.

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). Analysis of response-time distributions in the study of cognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 598–615.

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

Jacobs, A. M., & Grainger, J. (1992). Testing a semistochastic variant of the interactive activation model in different word recognition experiments. *Journal of Experimental Psychology: Human Perception and Performance, 4,* 1174–1188.

Kac, M. (1962). A note on learning in signal detection. *IRE Transactions on Information Theory, 8,* 126–128.

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

LaBerge, D. A. (1962). A recruitment theory of simple behavior. *Psychometrika, 27,* 375–396.

Laming, D. R. J. (1968). *Information theory of choice reaction time.* New York: Wiley.

Lee, W., & Janke, M. (1964). Categorizing externally distributed by stimulus samples for three continua. *Journal of Experimental Psychology, 68,* 376–382.

Link, S. W. (1975). The relative judgement theory of two-choice response time. *Journal of Mathematical Psychology, 12,* 114–135.

Link, S. W., & Heath, R. A. (1975). A sequential theory of psychological discrimination. *Psychometrika, 40,* 77–105.

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

Luce, R. D. (1986). *Response times.* New York: Oxford University Press.

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception and Psychophysics, 53,* 49–70.

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

McClelland, J. L. (1993). Toward a theory of information processing in graded, random, interactive networks. In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence and cognitive neuroscience* (pp. 655–688). Cambridge, MA: MIT Press.

McClelland, J. L., McNaughton, B., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex. Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102,* 419–457.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88,* 375–407.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 109–165). New York: Academic Press.

McCloskey, M., & Lindemann, A. M. (1992). Mathnet: Preliminary results from a distributed model of arithmetic fact retrieval. In J. I. D. Campbell (Ed.), *The nature and origin of mathematical skills* (pp. 365–409). Amsterdam: Elsevier.

McElree, B., & Dosher, B. A. (1993). Serial retrieval processes in the recovery of order information. *Journal of Experimental Psychology: General, 122,* 291–315.

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

Movellan, J. R., & McClelland, J. L. (1991). *Learning continuous probability distributions with the contrastive Hebbian algorithm* (Tech. Rep.

No. PDP.CNS.91.2). Pittsburgh, PA: Carnegie Mellon University, Department of Psychology.

Movellan, J. R., & McClelland, J. L. (1993). Learning continuous probability distributions with symmetric diffusion networks. *Cognitive Science, 17,* 463–496.

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

Muter, P. A. (1979). Response latencies in discriminations of recency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 5,* 160–169.

Nelder, J. A., & Mead, R. (1965). A simple method for function minimization. *Computer Journal, 7,* 308–313.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar based random walk model of speeded classification. *Psychological Review, 104,* 266–300.

Peterson, C., & Hartman, E. (1989). Explorations of the mean field theory learning algorithm. *Neural Networks, 2,* 475–494.

Pike, R. (1973). Response-latency models for signal detection. *Psychological Review, 80,* 53–68.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103,* 56–115.

Proctor, R. W. (1986). Response bias, criteria settings, and the fast–"same" phenomenon: A reply to Ratcliff. *Psychological Review, 93,* 473–477.

Proctor, R. W., & Rao, K. V. (1983). Evidence that the same–different disparity is not attributable to response bias. *Perception and Psychophysics, 34,* 72–76.

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 modelling accumulation of information when the rate of accumulation changes over time. *Journal of Mathematical Psychology, 21,* 178–184.

Ratcliff, R. (1981). A theory of order relations in perceptual matching. *Psychological Review, 88,* 552–572.

Ratcliff, R. (1985). Theoretical interpretations of speed and accuracy of positive and negative responses. *Psychological Review, 92,* 212–225.

Ratcliff, R. (1987). More on the speed and accuracy of positive and negative responses. *Psychological Review, 94,* 277–280.

Ratcliff, R. (1988). Continuous versus discrete information processing: Modeling the accumulation of partial information. *Psychological Review, 95,* 238–255.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review, 97,* 285–308.

Ratcliff, R., & Hacker, M. J. (1981). Speed and accuracy of same and different responses in perceptual matching. *Perception and Psychophysics, 30,* 303–307.

Ratcliff, R., & Hacker, M. J. (1982). On the misguided use of reaction time differences: A reply to Proctor and Rao (1982). *Perception and Psychophysics, 31,* 603–604.

Ratcliff, R., & McKoon, G. (1982). Speed and accuracy in the processing of false statements about semantic information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 16–36.

Ratcliff, R., & McKoon, G. (1989). Similarity information versus relational information: Differences in the time course of retrieval. *Cognitive Psychology, 21,* 139–155.

Ratcliff, R., & McKoon, G. (1997). A counter model for implicit priming in perceptual word identification. *Psychological Review, 104,* 319–343.

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

Ratcliff, R., & Rouder, J. F. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9,* 347–356.

Ratcliff, R., & Rouder, J. F. (in press). A diffusion model account of masking in two-choice letter identification. *Journal of Experimental Psychology: Human Perception and Performance.*

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96,* 523–568.

Smith, P. L. (1995). Psychophysically principled models of visual simple reaction time. *Psychological Review, 102,* 567–591.

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. (1969). The discovery of processing stages: Extensions of Donder's method. In W. G. Koster (Ed.), *Attention and performance: II* (pp. 276–315). Amsterdam: North-Holland.

Stone, M. (1960). Models for choice reaction time. *Psychometrika, 25,* 251–260.

Strayer, D. L., & Kramer, A. F. (1994a). Strategies and automaticity: I. Basic findings and conceptual framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 318–341.

Strayer, D. L., & Kramer, A. F. (1994b). Strategies and automaticity: II. Dynamic aspects of strategy adjustment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 342–365.

Swensson, R. G. (1972). The elusive tradeoff: Speed versus accuracy in visual discrimination tasks. *Perception and Psychophysics, 12,* 16–32.

Thomas, E. A. C. (1973). On a class of additive learning models: Error-correcting and probability matching. *Journal of Mathematical Psychology, 10,* 241–264.

Thomas, E. A. C. (1975). Criterion adjustment and probability matching. *Perception and Psychophysics, 18,* 158–162.

Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes,* Cambridge, England: Cambridge University Press.

Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review, 91,* 68–111.

Tuckwell, H. C. (1989). *Stochastic processes in the neurosciences.* Philadelphia: Society for Industrial and Applied Mathematics.

Usher, M., & McClelland, J. L. (1995). *On the time course of perceptual choice: A model based on principles of neural computation* (Tech. Rep. PDP.CNS.95.5). Pittsburgh, PA: Carnegie Mellon University, Department of Psychology.

Van Zandt, T., & Ratcliff, R. (1995). Statistical mimicking of reaction time distributions: Mixtures and parameter variability. *Psychonomic Bulletin and Review, 2,* 20–54.

Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. *Ergonomics, 13,* 37–58.

Vickers, D. (1979). *Decision processes in visual perception.* New York: 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.

Ward, L. M., & Lockhead, G. R. (1970). Sequential effects and memory in category judgments. *Journal of Experimental Psychology, 84,* 27–34.

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

Appendix

The Diffusion Model

The diffusion model is a continuous variant of the random walk model for choice reaction time (Laming, 1968; Link & Heath, 1975; Stone, 1960). It assumes that information is accumulated continuously over time and when the accumulated information reaches either one of the two response criteria, a response is generated. There are explicit mathematical solutions for a number of important features, for example, response probabilities, reaction-time distributions, and the distribution of processes that are not terminated at any point in time (see Ratcliff, 1988). This is a methodological advantage over other models that require simulation to produce predictions for these aspects of performance.

The equation for the diffusion process is a differential equation (the Fokker–Planck equation):

$$\frac{\partial}{\partial t}f(t,\,z) = \xi\frac{\partial}{\partial z}f(t,\,z) + \frac{s^2}{2}\frac{\partial^2}{\partial z^2}f(t,\,z),$$

where z is the starting point of the diffusion process, t is time, ξ is the drift rate, and s is the standard deviation in drift within a trial. With appropriate boundary conditions (see Ratcliff, 1978), the probability of a decision at boundary 0 (where 0 and a are the two boundary positions; see Figure 7) is given by

$$P(\xi) = (e^{-(2\xi a/s^2)} - e^{-(2\xi z/s^2)})/(e^{-(2\xi a/s^2)} - 1),$$

and the cumulative distribution of finishing times at the 0 boundary is given by

$$G(t,\,\xi) = P(\xi) - \frac{\pi s^2}{a^2}e^{-(z\xi/s^2)} \times \sum_{k=1}^{\infty}\frac{2k\,\sin(k\pi z/a)e^{-1/2(\xi^2/s^2+\pi^2k^2s^2/a^2)t}}{(\xi^2/s^2 + \pi^2k^2s^2/a^2)}.$$

To solve this equation, it is necessary to sum the infinite series numerically, but typically it converges by 40 or 50 terms.

To include variability in the drift ξ or variability in the starting point z, it is necessary to integrate the expressions for G and P over drift or starting point. This can be done numerically, and the expression used for this with a normal distribution for variability in drift ξ is

$$G(t,\,v) = \int_{-\infty}^{\infty} G(t,\,\xi)\,\frac{1}{\sqrt{2\pi\eta^2}}\,e^{-[(v-\xi)^2/2\eta^2]}d\xi.$$

Predictions from the diffusion model can also be produced by simulation. The simple random walk is used and limiting behavior is obtained by making the number of steps very large and the probability of a step to one boundary nearly equal to the probability of a step to the other boundary to conform to specific limits on these quantities (see Feller, 1968). In this article, the explicit solutions were used in modeling.

Fitting Reaction-Time Distributions and the Diffusion Model

For the diffusion model to fit reaction-time distributions, a summary of the empirical distributions is useful. The summary that has worked well in the past (Ratcliff, 1978, 1981, 1988) uses the ex-Gaussian distribution (the convolution of normal and exponential distributions) to give a summary of an empirical distribution (Heathcote, Popiel, & Mewhort, 1991; Hockley, 1984; Hohle, 1965; Ratcliff, 1978, 1979, 1981, 1988; Ratcliff & Murdock, 1976). The parameters of the ex-Gaussian distribution are the mean (μ) and standard deviation (σ) of the normal distribution that describes the leading edge of the reaction-time distribution and the mean of the exponential (τ) that describes the fall of the tail of the distribution. In practice, only the mean of the normal and the mean of the exponential (μ and τ) are used because they have been sufficient to fit the diffusion model accurately to data. The expression for the ex-Gaussian distribution is

$$g(t) = \frac{e^{-(t-\mu)/\tau+\sigma^2/(2\tau^2)}}{\tau\sqrt{2\pi}} \times \int_{-\infty}^{[(t-\mu)/\sigma-\sigma/\tau]} e^{(-y^2)/2}dy.$$

The process of obtaining a set of parameter values for the fit of the diffusion model to the data involved several steps. First, an initial set of values for the model parameters was picked (values for a, η, T_{er}, and v). Second, using the explicit equations of the diffusion model, the parameter values were used to generate predictions for the probabilities of "high" and "low" responses and predictions for the shapes of reaction-time distributions. Third, the ex-Gaussian parameters that best described the predicted reaction-time distributions were determined. Fourth, the predicted values of μ, τ, and response probability were each subtracted from the corresponding values for the data, and these differences were squared and the sum of these over three different conditions was used as the function to be minimized by the SIMPLEX minimization routine (Nelder & Mead, 1965). SIMPLEX adjusted the model parameters a, η, T_{er}, and a value of v for each condition to produce a minimum sum of squares for the function

$$(\mu_{cex} - \mu_{cth})^2 + (\tau_{cex} - \tau_{cth})^2 + (\mu_{eex} - \mu_{eth})^2 + (\tau_{eex} - \tau_{the})^2 + (p_{ex} - p_{th})^2,$$

where *ex* is experimental, *th* is theoretical, *cex* is correct experimental, *cth* is correct theoretical, *eex* is error experimental, *eth* is error theoretical, and *p* is the probability of a response. In more recent work, we fit the cumulative distributions directly (eliminating the ex-Gaussian intermediate step) and the results are nearly identical.

Qualitative Behavior of the Model

Here we describe how the diffusion model accounts for a number of standard phenomena in the reaction-time domain. First, the model accounts for the speed–accuracy tradeoffs obtained when subjects vary speed at the expense of accuracy by varying response boundary positions. When the boundaries are close to the starting point, processes reach the boundaries quickly, but they can hit the wrong boundary by mistake, leading to errors. When the boundaries are moved further apart, the time to hit a boundary is increased, and processes that would have hit the wrong boundary by mistake when the boundaries were close to the starting point now have room to correct themselves, leading to fewer errors.

Second, the distributions of response times are skewed to the right. This occurs through the geometrical form of the process. Equal increments in drift rate (in the vertical direction) map into increasing increments on the decision boundary as time increases. Thus, differences in the fastest processes are small, whereas differences among the slowest processes are large. This also leads to predictions of statistical interactions in reaction time from equal size differences in drift rate.

Third, the accuracy of processes that have not yet reached a response criterion was studied by Meyer et al. (1988). They found that the accuracy of nonterminated processes rose rapidly to a low asymptote and then remained constant over a relatively long time interval. This seemed inconsistent with continuous models that appeared to predict a gradual rise in accuracy of nonterminated processes as a function of time. In fact, the diffusion model and several other sequential sample models (counter and runs models) predicted the same effects as found in the experimental data (counterintuitively), and the data provided strong confirmation for that class of models.

Fourth, speed–accuracy functions from the response signal procedure can be modeled in two different ways. Ratcliff (1978) assumed that the response boundaries were placed far from the starting point and that when the signal to respond was presented, the position of the process was determined and the response was given that corresponded to the position relative to the starting point. This assumption resulted in a simple formula for response signal functions,

$$d' = \frac{d'_{ASY}}{\sqrt{1 + s^2/\eta^2 t}}.$$

This function is a reasonable competitor for the commonly used exponential approach to a limit that is used to summarize response signal functions, and a large number of observations are needed to distinguish between the two. Ratcliff (1980) provided expressions for situations in which the rate of accumulation of information changes during the time course of processing (e.g., Dosher, 1984; Gronlund & Ratcliff, 1989; Ratcliff & McKoon, 1982, 1989).

The second way to model response signal functions assumes that response signal functions are a mixture of processes that have terminated at a response boundary (and, in some cases, are being held up until the signal to respond is given) and processes that have not terminated as measured using the Meyer et al. (1988) method (see Ratcliff, 1988). A mixture of these two sources of responses shows an increasing response signal function that mimics the typical response signal functions.

Equations for the GRAIN Model and the Mean Field Learning Algorithm

The GRAIN model was designed to introduce variability into processing, to use this variability to account for reaction-time data, and to provide a stochastic mechanism for generating error responses. Figure 16 shows the three-layer network used in simulations in this article, and each node at one level is connected to all the nodes at levels above and below. At test, an input pattern is presented to the input layer and activation flows through the network to provide activation at the output layer. Each node in the network has an activation value that ranges from +1 to −1. For a node at level *i*, the net input from the level below it (level *j*) is given by

$$net_i(t) = \sum_j w_{ij}a_j(t) + N(0, \sigma) + bias_i,$$

where $N(0, \sigma)$ is a random number from a normal distribution with standard deviation σ, w_{i_j} are weights connecting nodes *i* and *j*, and a_j are the activation values at level *j*. Bias can be thought of as an extra connection or weight from a node that always has activation 1. A running average of the net input is computed (*anet*), and then the average net input is converted to activation using a nonlinear transformation (the *tanh*):

$$anet_i(t) = \lambda net_i(t) + (1 - \lambda)anet_i(t - 1)$$

$$a_i(t) = tanh\,[anet_i(t) + bias],$$

where the *tanh* function is

$$tanh\,(x) = (e^x - e^{-x})/(e^x + e^{-x}).$$

For the model presented in this article, there is one output node, activation is examined in this node after every update, and the process is terminated and a response produced when the activation becomes larger than +0.9 or smaller than −0.9.

Mean Field Learning Algorithm

The mean field learning algorithm is a discrete version of the Boltzman learning algorithm (Peterson & Hartman, 1989). It involves two phases of processing, one in which the outputs are fixed at their desired values and another in which they are allowed to vary freely. In both phases, activation is allowed to flow in both directions (in a multilayer network) and activation is allowed to settle down gradually until the system reaches an asymptotic state (using an annealing schedule). Once activation has settled down, the weights are altered and the next phase of processing begins. The basic idea is that the clamped phase (when the outputs of the network are fixed at the desired target values) produces activation values in the hidden layer, and in the free state, activations differ by some amount from those values. The weights are modified to bring the free state activations closer to the clamped state activations. When this process is complete, the outputs in the free state will be the same as the outputs in the clamped state (both in the hidden layer and in the output layer), which means the learning algorithm will have trained the network to produce the desired outputs.

Consider the network in Figure 16. Input activations are denoted w_i, hidden layer activations are a_j, output layer activations are b_k, weights between input and hidden layers are s_{ij}, and weights between the hidden layer and output layer are r_{jk}. In the free phase, activation at the hidden layer is given by

$$a_j^f = tanh\left(\frac{1}{\tau}\sum_i s_{ij}w_i + \frac{1}{\tau}\sum_k r_{jk}b_k^f\right),$$

and activation at the output layer is given by

$$b_k^f = tanh\left(\frac{1}{\tau}\sum_j r_{jk}a_j^f\right),$$

where the superscript *f* refers to the free phase of processing. In the clamped phase (where the outputs are set to the target values)

$$a_j^c = tanh\left(\frac{1}{\tau}\sum_i s_{ij}w_i + \frac{1}{\tau}\sum_k r_{jk}b_k^c\right),$$

where the superscript *c* refers to the clamped phase. To implement the noise and running average assumptions from GRAIN in the mean field algorithm, the three equations above have noise added and a running average of activation computed.

The process is iterative for the free phase (the activations for *a* are computed, then *b*, then *a*, etc.), with τ (a scaling parameter) being reduced to about one third of its initial value by 100 iterations (simulated annealing). Once the activation values for the free and clamped phases for all the nodes are determined, the weights are modified by the equations

$$\Delta r_{jk} = \eta(a_j^c b_k^c - a_j^f b_k^f)$$

and

$$\Delta s_{ij} = \eta(w_i a_j^c - w_i a_j^f) = \eta w_i(a_j^c - a_j^f).$$

Received June 26, 1995
Revision received April 27, 1998
Accepted June 1, 1998 ■