

PSYCHONOMIC SOCIETY KEYNOTE ADDRESS

Traps in the route to models of memory and decision

W. K. ESTES

Indiana University, Bloomington, Indiana

It is proposed that the products of investigations of learning, memory, and decision over the last half century that are most likely to endure have resulted from interactions between models and experimental research. In this article, some of the traps that must be coped with to make fruitful interactions possible are examined and illustrated with case studies from research on probability learning, category learning, and recognition memory. Topics addressed include functions of models in research; the logic of model testing; fitting models to signal plus noise; values and hazards of averaging data; and potential contributions of neural science to the development of cognitive models.

Over more than a half century of experience in research on learning, memory, and decision, I have come to believe that the most substantial and enduring advances have not been in the accumulation of empirical facts or the construction of models, but in the production of fruitful interactions between models and experimental research. Most experimental facts require continual reinterpretation and most models drop by the wayside like autumn leaves, but the results of interactions between models and experiments constitute most of our generalizable knowledge.

Success in the interactive research effort depends not only on clearly formulated models and well-conducted experiments, but, just as importantly, on sound interpretations of the results of applying the models to the experiments. This interpretive phase of the effort is in some respects the most difficult, and I take as my main task in this article an account of some of the issues that have to be resolved and some of the traps that have to be avoided in order for the process to run to a successful conclusion. As a preliminary, I turn to a review of the basic concept of applying a model to data as it has evolved since its first rudimentary instantiation in the literature of memory and decision more than a century ago.

Applying Models to Experiments

Details of techniques for fitting curves, or, more broadly, formal models, whether mathematical or computer imple-

mented, to data are available in many sources (e.g., Bush, 1963; Wickens, 1982). Here, I wish only to bring up advances in methodology that are of particular importance in the context of this article.

A pioneering exercise in curve fitting. The earliest instance of fitting a mathematical function to psychological data that I am aware of is an exercise carried out by Hermann Ebbinghaus, the founder of the experimental psychology of memory. In 1879, Ebbinghaus conducted a study of the learning and retention of lists of artificial words (“nonsense syllables”), using himself as the experimental subject. Each of some 150 experimental episodes constituted the learning of several lists of syllables followed after an interval by a memory test. Amount remembered was found to decline over time between learning and testing in a reasonably orderly fashion, but with some fluctuations. Ebbinghaus noted that, because of the fluctuations, the empirical curve could not constitute a law of retention, but he reasoned that a mathematical function made to pass through the empirical points on a graph might reveal a lawful underlying trend.

He evidently considered a number of functions and discovered that his retention data were described quite satisfactorily by the function

$$R = \frac{100k}{(\log t)^c + k}, \quad (1)$$

where R denotes amount remembered, k and c are constants, and $\log t$ is the natural log of time in minutes between learning and testing. This result is illustrated in Figure 1, which I constructed from the tabulation of observed and theoretical values of R reported by Ebbinghaus (1885/1964, p. 78). Ebbinghaus must have determined the values of the constants (“free parameters”) k and c that would produce this fit by simple trial and error, because no other method was available to him.

This article presents in substance the author's Governing Board Keynote Address to the 42nd Annual Meeting of the Psychonomic Society, Orlando, FL, November 15, 2001. Preparation of the article was supported in part by NSF Grant SBR 96-10048. I am indebted to John R. Anderson, James L. McClelland, and John T. Wixted for extremely useful and timely comments on the original draft and to Kay Estes for tireless editorial assistance. Reprint requests or correspondence about this article should be addressed to W. K. Estes, Department of Psychology, Indiana University, Bloomington, IN 47405-7007 (e-mail: wkestes@indiana.edu).

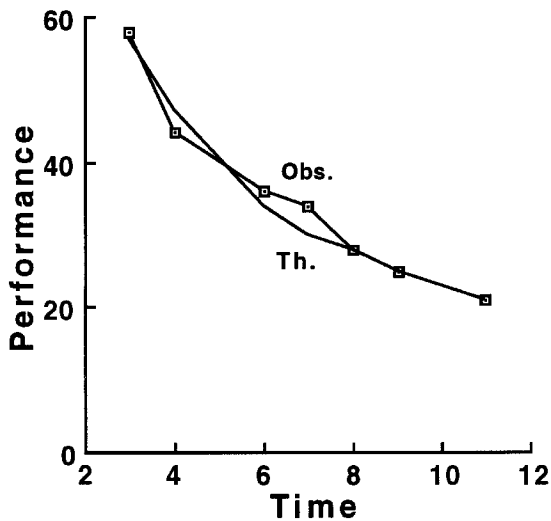


Figure 1. The first instance of fitting a mathematical function to an observed retention curve: $R = 100k/[\log t)c + k]$, where R = amount remembered, t = natural log of retention interval in minutes, and k and c are constants. Obs., observed values. Th., theoretical values. Data are from Ebbinghaus (1885/1964, p. 78).

Estimating parameters of models. Although Ebbinghaus's little tour de force did not spark immediate interest among investigators of learning and memory (not a large constituency at the time), the idea of fitting mathematical functions to empirical curves reemerged several decades later and sparked a new tradition in which textbooks and handbook chapters on human learning and memory had to be salted with chapters or sections on somewhat mechanical curve fitting (e.g., Hilgard, 1951; McGeoch, 1942). The chore of selecting appropriate values of the free parameters in functions describing learning or retention was based on algebraic heuristics often requiring a rare combination of technical virtuosity and tenacity (Bush, 1963). With the advent of more complex models of cognitive processes in the 1960s, however, these heuristics for parameter estimation, no matter how tirelessly applied, were no longer adequate, and a new approach was essential.

The new approach was at hand, owing to developments outside of the field, and it meant returning to the trial-and-error method of estimation used by Ebbinghaus. The revival of Ebbinghaus's method was implemented in computer programs of a type dubbed "hill climbing." These programs search the space of all possible values of a model's free parameters for the set that yields the best fit to data as measured by the minimum sum of squared differences between observed and theoretical performance measures (or the maximum likelihood of the data given the fitted model).

With the steeply increasing power of even desktop computers over the past decade, it is now possible for an investigator to fit the kinds of models associated with current research on memory and decision to individual data even for large groups of subjects. The importance of this development will become evident in the next section.

A new look in parameter estimation. Paralleling the advances in computational methodology was a similarly important deepening of experimental psychologists' understanding of the purposes that can be served by applying models to data. In Ebbinghaus's equation for retention, the parameters were merely numerical constants chosen so as to make the mathematically defined curve pass through points on a graph representing empirical values and had no other significance. For present day investigators, in contrast, estimation of parameters is a critical step in moving from descriptions of phenomena in ordinary language to representations in a theoretical plane. This kind of translation was familiar at a verbal level to the Gestalt psychologists of the 1920s who wrote of a "psychological space" that served as the locus of mental activities (Koffka, 1924) but did not have the quantitative tools to go beyond pictorial analogy. Outside of the Gestalt movement, experimental investigators of cognition during the next quarter century must have been aware of the need for such a concept, in view of widely spreading knowledge of how stimulus information impinging on sensory receptors is transduced into quite different forms, with the outputs of the transducers providing the basis for cognitive processes.

The prevailing strategy, however, was to limit experiments to simple situations in which dimensions appropriate for describing higher order representations might be presumed to differ little from the physical dimensions customarily used in describing experimental situations. The cost of this strategy of simplification was to limit theoretical efforts largely to searches for very limited empirical "laws" and thus to severely curb progress toward functional theoretical models.

A breakthrough eventually appeared in the form of a quantitative methodology for relating stimulus similarity, a central concept in theories of memory and classification, to distance in a hypothesized cognitive space where stimulus representations are manipulable by cognitive operations. The prime movers of this development were, first, with respect to basic memory theory, Roger Shepard and his associates (Shepard, 1958, 1974; Shepard, Romney, & Nerlove, 1972), and later, with respect to classification and learning, Robert Nosofsky (Nosofsky, 1984, 1992). As a consequence of this work, one now can translate an observed relation between performance in a task and physically described stimulus variables into a relation in a conceptual space by estimating the appropriate parameter in a transformation function. Thus, parameter estimation has evolved from being simply a tool for curve fitting to being a broadly applicable method for increasing the theoretical informativeness of experimental data.

In this article, I use selected case histories to show how interactions of models with experiments generate new knowledge, for the process is difficult to grasp apart from actual research. The plan in any instance is to apply a model to a set of research results in order to provide a framework for theoretically relevant analyses. Success of the plan depends critically on the proposed model being appropriate

to the given situation, so effectuation of the plan must begin with testing of the model. The definition of “appropriate” as used in this context needs explication, however, as do the principles that guide effective model testing. Thus, I shall set the stage for these case studies by reviewing the logic of model testing. To avoid encountering all of the complexities of the process at once, I first view model testing as a logical exercise, then consider how central concepts change when we move from the realm of abstract logic to the realm of scientific research.

The Logic of Model Testing

What do we mean by a correct (appropriate) model for an assemblage of data? The generally accepted criterion, I believe, is that the model is necessary and sufficient for prediction of the data. However, the notions of necessity and sufficiency need close scrutiny, starting with a shift from the informally stated criterion to one given in standard logical notation (Suppes, 1957). Corresponding to the notion of a model predicting data is a relation of implication conventionally expressed as

$$P \rightarrow Q,$$

read as “*P* implies *Q*,” where *P* and *Q* are statements and \rightarrow is the symbol for implication. Reference to logical truth tables tells us that the implication is false if *Q* is false when *P* is true; otherwise the implication is true. Importantly for our present purposes, an alternative reading of the expression is “*P* is a sufficient condition for *Q*.” To capture necessity of a model for prediction of the data, we need another implication

$$\neg P \rightarrow \neg Q,$$

read as “not *P* implies not *Q*,” the implication being false if *Q* is true when *P* is not true. The alternative reading of present interest is “*P* is a necessary condition for *Q*.”

To translate the formal notation into terms suitable for a discussion of models in research, I shall take *P* to be a description of a model and *Q* a description of the data of a test situation. In these terms, it can be concluded that if a model generates predictions about a test situation and the predictions prove to be incorrect, then the model is not confirmed (and must be judged inappropriate for the situation). If the predictions prove to be correct, however, all we know is that the assumptions of the model are sufficient for prediction of the data. Only if we find that changing the assumptions of the model (changing *P* to $\neg P$) produces an incorrect prediction of the data (changing *Q* to $\neg Q$) can we conclude that the assumptions are necessary for prediction of the data and that the model satisfies the criterion of appropriateness given above.

It must be understood that formal logic is not a set of cookbook rules for model testing. The role of logic is to provide general principles that should be taken into account, explicitly or implicitly, at every step in the evaluation of models and their interactions with experiments, as will be illustrated in the following sections.

Empirical Evidence in Model Testing

The error component of behavioral models. In the setting of this article, we are not concerned primarily with pure logic, where correctness of a proposition can be determined by a formal proof, but with science, where correctness depends on empirical evidence. Reasoning about relations between models and evidence must be logical, but it must also take account of some shifts in the meanings of concepts when we move from the logic of sentential inference to the logic of scientific model testing. One of these has to do with the basic conception of predicting or describing data by a model.

When the logical expression $P \rightarrow Q$ is brought into the realm of model testing, *Q* is initially translated as “a description of test data.” The translation does not, however, make explicit a crucial distinction between “description of data” as it is interpreted in everyday life and in theoretically oriented science. This distinction derives from the universally accepted assumption that every empirical measure includes some kind of error of measurement. In psychological test theory, for example, a person’s score on a test is assumed to be a combination of the person’s true score and an error component:

$$\text{Test Score} = \text{True Score} + \text{Error}.$$

Because error is defined as a normally distributed variable with a mean of zero, averaging a person’s scores on a number of items can be assumed to reduce the error and yield an improved approximation to the true score.

An analogous argument applies to experimental research in a way familiar to everyone who has used an analysis of variance. The response measures that a subject produces are assumed to reflect the subject’s true responses to experimental variables plus error components:

$$\text{Performance} = \text{True Response} + \text{Error}.$$

The goal of fitting a model to data is to determine whether the model describes (or predicts) the true response. However, this goal cannot be reached directly, for a model fitted to data actually describes a combination of true response and error (or, in the terminology of signal detection theory, a mixture of signal and noise). To achieve the goal of model fitting, it is necessary somehow to filter out the error component of the performance measure, usually by some kind of averaging.

Coping With Error

The two faces of averaging. Confronted with the problem of applying models to data that comprise true patterns plus error, we have to rely mainly on some kind of averaging to reduce the contribution of error. However, although averaging data can help bring out real trends, it is important to recognize that it can also be a source of distortions. This point can be explicated in terms of a generic expression for the relation between a measure of performance and its equivalent in a model: Suppose that an experiment is conducted with a group of subjects and that the true response

pattern for individual i on trial (or time interval) n can be described by the function $f(i, n, \theta_{i1}, \theta_{i2}, \dots)$, where $\theta_{i1}, \theta_{i2}, \dots$ are values of parameters for that subject. Then $P_{i,n}$, a measure of the individual's performance, can be expressed in the form

$$P_{i,n} = f(n, \theta_{i1}, \theta_{i2}, \dots) + e_i, \quad (2)$$

where e_i is an error term associated with subject i , assumed to be normally distributed with a mean of zero.

It will be convenient to begin an analysis of the problems solved and the new problems raised when data are averaged by assuming that the performance measure in Equation 2 comes from a research episode constituting an application of a new model to a single set of experimental results for the purpose of determining whether the model correctly predicts the true response pattern.

Ideally, one would like to fit the model to the data of an individual subject, but to obtain stable data there must be replication within the experiment. Replication with the same subject is feasible when performance is at a steady state (as in many situations that arise in psychophysics and in the test series of category learning or study–test recognition experiments). In this case, an average of a subject's performance score, $P_{i,n}$, over replications can be represented as

$$M[P_{i,n}] = M[f(n, \theta_{i1}, \theta_{i2}, \dots)] + M[e_i], \quad (3)$$

where, for any quantity x , $M[x]$ denotes the mean of x over replications. As the number of replications becomes large, the error term becomes negligibly small, leaving in the limit

$$M[P_{i,n}] = M[f(n, \theta_{i1}, \theta_{i2}, \dots)]. \quad (4)$$

One caveat must be emphasized: Averaging will bring the mean value of $P_{i,n}$ closer to the true value for a given subject only if the term $f(n, \theta_{i1}, \theta_{i2}, \dots)$ in Equation 2 represents the correct (most appropriate) model for the situation. Thus, the effects of statistically smoothing the data cannot be dissociated from the problem of judging when one has arrived at an appropriate model.

When performance is not stationary, as with learning or retention curves, some reduction of the error component of $P_{i,n}$ can be achieved by averaging over successive blocks of trials. The contribution of error is similarly reduced, provided that the values of the parameters remain constant over trials. However, in practice, one cannot use a large enough trial block for $M[e_i]$ to approach zero without sacrificing virtually all information about the form of the function.

A new problem arises in the much more common cases when individual data are too unstable for useful model fitting and one is led to average the data over subjects before applying a model. Two questions arise:

1. Will computing an average at successive values of n for a group of subjects who differ with respect to values of the parameters produce a group curve that is representative of the curves for individuals?

2. Will fitting a model to averaged data for a group yield evidence about the appropriateness of the model for predicting the behavior of individuals?

Addressing both questions requires averaging both sides of Equation 2 over i , which yields the expression

$$M[P_n] = M[f(n, \theta_1, \theta_2, \dots)] + M[e] \quad (5)$$

for the group mean. The error term, $M[e]$, tends to zero as the number of subjects becomes large. However, the values of the parameters will, in general, vary from subject to subject, and therefore, except under special circumstances, the group mean will not be a function of the same form as that in Equation 5 with mean values of the parameters. That is, $M[f(n, \theta_1, \theta_2, \dots)]$ is not equal to $f(n, M[\theta_1], M[\theta_2], \dots)$, where $M[\theta_1]$ and so forth denote mean parameter values for the group.

Experience with hazards of averaging. Understanding of the costs associated with applying models to group data has come slowly. A trap that lurks in the background was identified in the 1950s: A number of investigators raised alarms about dangers of averaging data—in particular, that an average curve may not be representative of the trends for the individual subjects whose data are averaged (Bakan, 1954; Sidman, 1952)—and I published an initial attempt at showing how artifacts of averaging might be coped with, explicating, for example, conditions under which group means do or do not preserve the forms of individual functions (Estes, 1956).¹ It is not easy, however, to change the habits of people who are comfortable with traditional ways of doing things, and developers of cognitive models have continued to rely for support mainly on the fitting of functions such as curves of learning, retention, and generalization to averaged data.

An extremely visible case is work on curves of practice and retention. An extensive review of literature on performance of motor and cognitive skills by Newell and Rosenbloom (1981) produced the generalization that curves of practice are described better by power functions than by exponential functions. This generalization was widely accepted, and the power function took on a basic role in models of skill acquisition. Concurrently, a similar scenario played out with respect to retention of skills, where the power function (which became the “power law”) was the preferred descriptor of retention curves.²

However, this entire development depended on the fitting of curves to group data, averaged over subjects. The latest chapter in this story is a flurry of new studies in which investigators have analyzed artificial data generated by computer programs based on either power or exponential functions (R. B. Anderson & Tweney, 1997; Heathcote, Brown, & Mewhort, 2000). These studies have shown that artifacts of averaging can be large for these functions, and that, especially for practice curves, there is a pervasive tendency for averaged data to be fit better by power functions even when performance of the individuals is known to conform to exponential functions. This tendency is illustrated in Figure 2, which shows a power function in comparison with an exponential function in the upper panel and the same power function in comparison with the average of three exponential functions in the lower panel.³

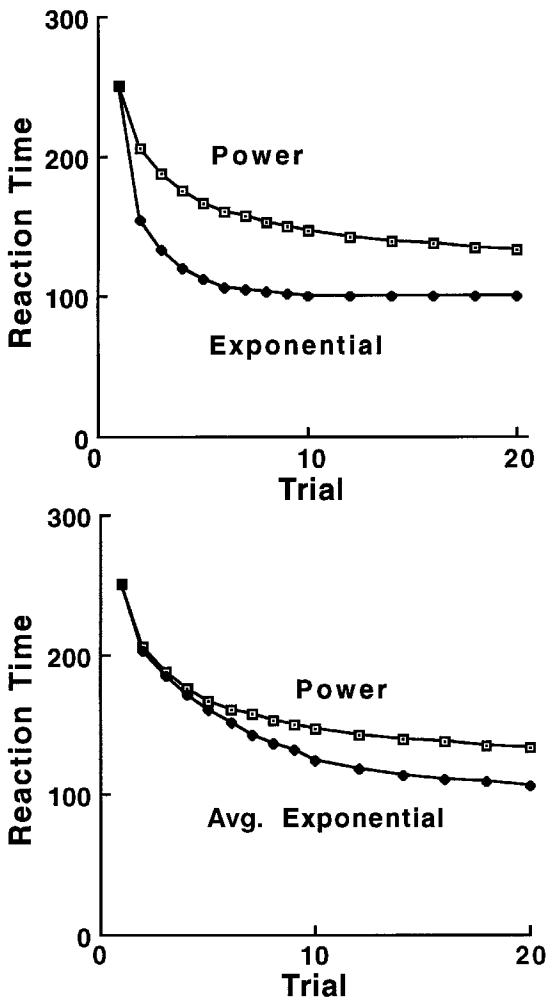


Figure 2. Comparison of a power with an exponential function (upper panel) and of the same power function with the average of three exponentials, differing only in value of the rate constant (lower panel).

The magnitudes of distortions by averaging depend on many factors, including noisiness of the data, variability of parameter values over individuals, and ranges of independent variables. Plainly, conclusions about the forms of functions that describe individual behavior must be justified by meticulous analyses of susceptibility to distortions whenever averaged data are used in tests of models. An analysis of this kind has been accomplished by Wixted and Ebbesen (1997) for retention curves and indicates that, for some conditions at least, retention is described better by power functions than by exponential functions. To my knowledge, no comparable analysis has been reported for curves of practice.

On the Selection of Appropriate Models

The criterion problem. Assessment of the appropriateness of a model for an empirical domain cannot be ac-

complished by testing the model in isolation, for there are no absolute definitions of success or failure of empirical tests. In practice, a minimal criterion for success is defined in terms of a significance level in a statistical test of the fit of the model to data. However, no way has been found to achieve general agreement on the required significance level, and, in any event, meeting some fixed criterion is not enough for scientific purposes, because it bears only on sufficiency of the model for predicting the data. In order to determine whether the assumptions of the given model are necessary for successful prediction, some kind of comparative testing is needed. In practice, investigators have turned to evaluations of relative success by comparing a model with alternative models on each of a succession of tests. There is no strict criterion for terminating the process, but if on any test the reference model generates the less accurate account of the test data, it is disconfirmed; otherwise, it is retained as the provisionally correct model.

Model complexity. This relative testing strategy appears sensible on the surface, but below the surface it encounters serious problems. For any research situation, the alternative models have usually been developed by different investigators and differ from the reference model in numerous respects. Perhaps the most important of these respects is model complexity. Since long before the pioneering work of Ebbinghaus on the psychology of memory, a driving motive behind the search for mathematical accounts of data throughout the sciences has been to find the simplest model that can fit a given set of data. Until very recently, however, evaluations of simplicity, or complexity, have been based only on subjective judgments, making it impossible in many cases to identify objectively the “simplest model” among any set of alternatives. This frustrating situation changed dramatically in the 1990s as a consequence of advances in the formal treatment and measurement of model complexity (Grunwald, 2000; Myung, 2000).⁴ Owing to this work, appraising the complexity of a model no longer depends only on counting its free parameters, but can also incorporate techniques that deal quantitatively with structural complexity; as a consequence, it has become possible to compare a model only with competitors of similar complexity.

Unfortunately, the same line of research that opened the way to measuring model complexity uncovered a new problem: As the models involved in a comparison become increasingly complex, it becomes increasingly likely that the model selected by the relative testing strategy will merely be the one that has happened to be favored by the particular sampling of the error constituent of the data (Myung, 2000). Thus, the relative testing strategy can be expected to be successful only when the models involved are relatively simple (roughly speaking, having no more than three or four free parameters to be estimated from data).⁵

Even when relative testing of complex models happens to hit on the best solution, the result is usually uninformative, for it is impossible to identify the particular processes or parameters that were responsible for the suc-

cess. Thus, an improved strategy is, at each step, to compare a reference model with an alternative version of the same model that differs from it with respect only to inclusion or exclusion of a single parameter or process. One gains information, not only about the sufficiency of the reference model for predicting the test data, but also about the contribution of the component whose exclusion leads to (relative) disconfirmation of the model. Some products of this “selective testing” strategy, as it has been implemented in research on memory and decision, will be identified in the case studies discussed in the following sections.

Case Study I: Probability Learning

I am occasionally asked whether probability learning has not become an obsolete topic in psychology. It seems scarcely credible to me that the answer could be “yes,” because there can be few cognitive activities as pervasive and adaptive in ordinary life as assessing the probabilities of alternative outcomes of choices (Gallistel, 1990; C. Peterson & Beach, 1967; Pitz, 1980). What may be obsolete is the issue of “matching versus maximizing”—that is, learning to match choice probabilities to outcome probabilities versus learning to make choices that maximize long-term success rates. That issue has not yielded to any simple and general resolution and now is rarely seen in the titles of research articles. However, a process of probability learning frequently appears as a component of models of categorization (Ashby & Alphonso-Reese, 1995), psychophysical judgment (Atkinson, Carterette, & Kinchla, 1962; Dorfman, 1969), decision making (Parks, 1966; Slovic, Lichtenstein, & Fischhoff, 1988), and two-person interactions (Burke, 1959; Estes, 1960; Estes & Suppes, 1959; Suppes & Atkinson, 1960).

The basic paradigm: Predictive responses and non-contingent event probabilities. This case history has to do with a phenomenon first discovered in the 1930s and known at the time (quaintly, it now seems) as “verbal conditioning” but later as “probability learning.” I first encountered this phenomenon in a research report by Grant, Hake, and Hornseth (1951). These investigators combined a conception of learning environmental probabilities due to Brunswik (1939) with an experimental procedure developed by L. G. Humphreys (1939) and conducted an experiment in which subjects had the task of predicting on each trial whether a ready signal would or would not be followed by a flash of a light. For different groups, the light occurred with different fixed probabilities, independently of the subjects’ responses. Although the learning series was only 60 trials in length, the data suggested that subjects’ probabilities of predicting that the light would appear were tending, on the average, toward asymptotes that would match the true probabilities.

The linear model for probability learning. I was not excited by this finding at the time, but recalled it some years later in connection with some mathematical explorations I had been doing on possible extensions of a statistical model for simple associative learning (Estes, 1950). In this model,

changes in mean response probability on reinforced and unreinforced trials were expressible in the functions

$$p_{n+1} = p_n + \theta(1 - p_n) \quad (5)$$

and

$$p_{n+1} = (1 - \theta)p_n, \quad (6)$$

respectively, where p_n denotes mean response probability on trial n , and θ is a “learning rate” parameter. Because of the simple form of these equations, this case of the statistical model has come to be known as the “linear model” (an appellation coined by Bush & Mosteller, 1955, who arrived at the same equations from different assumptions).

Reflecting on the results of the Grant et al. (1951) study, I observed that if Equation 5 applied on each reinforced and Equation 6 on each nonreinforced trial of a series, it would be predicted that a theoretical learning curve would follow a negatively accelerated course approaching the true reinforcement probability asymptotically.

An experiment designed to test this implication of the model was reported by Estes and Straughan (1956). The results appeared generally confirmatory: (1) Predicted learning curves for several conditions fitted the observed curves quite closely, in each case approaching probability matching in later trial blocks; and (2) with the rate parameter, θ , estimated from data of an initial series, a priori predictions of curves for trials following a change in event probabilities were encouragingly successful.

Having been sensitized to the hazards of averaging data, I examined the data of Estes and Straughan (1956) a few years later (Estes, 1964) to see whether the apparent support for the linear model could reflect distortions by averaging. Formal methods adequate for the task being lacking at that time, I had recourse simply to a compari-

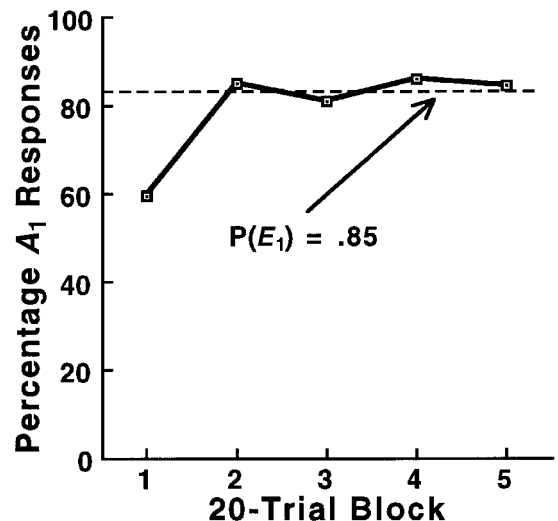


Figure 3. Mean learning curve for a sample of 8 subjects from the study of Estes and Straughan (1956).

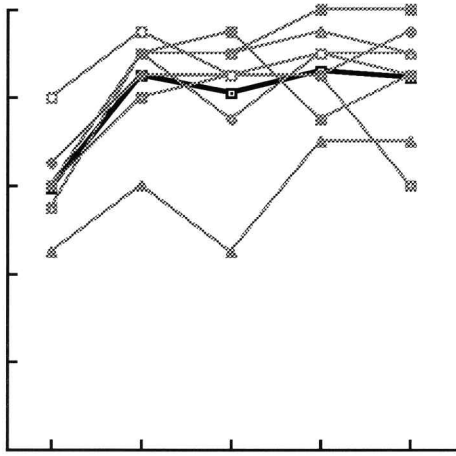


Figure 4. Mean curve from Figure 1 reproduced together with the curves for the 8 individual subjects.

son of individual and group trends. For this purpose, I used a sample of data consisting of the learning curves for the first 2 subjects in each of four conditions of the study that included a series of trials on which the events E_1 and E_2 that the subjects were trying to predict occurred with probabilities .85 and .15, respectively, independently of the subjects' responses. The mean learning curve for this sample, shown in Figure 3, clearly exhibits the properties implied by the model—a negatively accelerated approach to an apparent asymptote in the vicinity of .85, the probability matching value.

In Figure 4, the mean curve is reproduced and the curves for the 8 individual subjects are added. There were individual differences in response levels, but on the whole, the trend for the mean seemed quite representative of the individual trends. To give additional perspective on this result, I computed simulations of the model for 8 hypothetical subjects, and the mean curve obtained is shown, together with the eight individual curves, in Figure 5. The range of individual differences in trends for the 8 simulated subjects appeared similar enough to that for the real subjects shown in Figure 4 to support the conclusion that the predictions of mean learning curves in the Estes and Straughan (1956) study had not suffered material distortions due to averaging.

An alternative to the linear model: The pattern model. With these results in hand, was I in a position to conclude that the demonstrated ability of the linear model to predict mean learning curves could be taken as a strong confirmation of the model? The answer is “no.” I had taken out some insurance against hazards of averaging, but still must attend to the demands of logical reasoning. The model had survived a fairly comprehensive test without disconfirmation, but the impressive accounts of data did not warrant a conclusion that the model was correct. Thus, I turned to another of the routes that might yield evidence about correctness of the model—discovering whether

other models of comparable complexity predicted curves of learning as well as or better than the linear model.

Casting a wide net for such models, I did indeed uncover a plausible candidate, which came to be known as the “pattern model” (Estes, 1959). In this model, it is assumed that each stimulus situation that can arise at the onset of a trial in a probability learning experiment is perceived and remembered as a unitary pattern, which at any time is associated with exactly one of the alternative trial outcomes, E_1 or E_2 .⁶ The course of events on any trial, n , of an experiment is schematized in Figure 6.

When the stimulus pattern is associated with E_1 on any trial, the subject predicts E_1 , and when the pattern is associated with E_2 , the subject predicts E_2 . Learning occurs only on errors. When a prediction of E_1 occurs and is incorrect (i.e., E_2 occurs), then with probability c the current stimulus pattern becomes associated with E_2 . Similarly, if a prediction of E_2 occurs and is incorrect, then with probability c the current pattern becomes associated with E_1 . The “learning parameter” c is the one constant in the model that must be estimated from some aspect of the data of an experiment before a full quantitative account of the performance statistics can be generated.

Interpretation of the pattern model as rule selection.

It may be noted that although the pattern model emerged within the framework of associative learning theory, it might as aptly be classified as a rule selection (or hypothesis testing) model. In the simplest case of the model, the same stimulus pattern, S , occurs on every trial, and a subject has just two rules available:

When S occurs, predict event E_1 ,

and

when S occurs, predict event E_2 .

On the first trial of an experiment, a rule is chosen by the subject at random and continues in use until it yields an in-

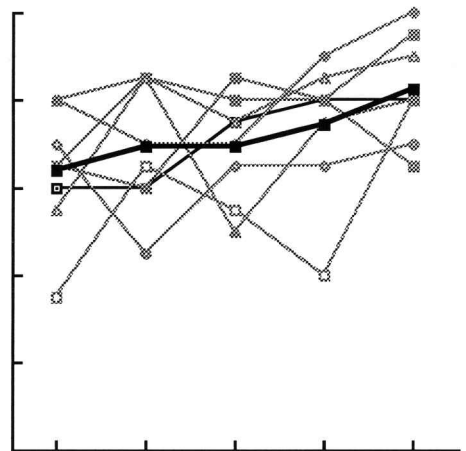


Figure 5. Results of a computer simulation of 8 hypothetical subjects who learn in accord with the linear model.

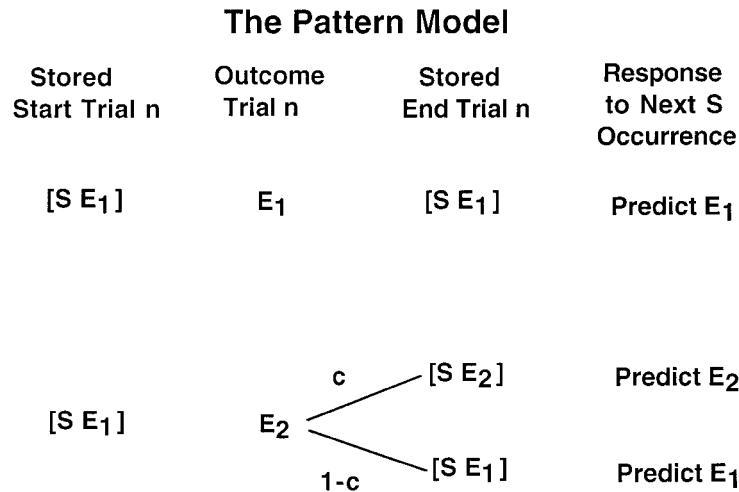


Figure 6. Assumptions of the pattern model for an experiment in which each trial begins with the same stimulus pattern S . A representation of S is stored in memory together with trial outcome E_1 at the start of trial n , and this representation changes or does not change on trial n in accordance with the outcome of the trial.

correct prediction; then, with probability c , the alternative rule is chosen and continues in use until it, in turn, yields an incorrect prediction, and so on. Although the verbal label for the model has changed, the computational machinery has not, and therefore the same sequence of responses is predicted whether the model is viewed as associatively based or rule based.

A decisive test of the linear versus pattern models.

An important “theorem” for the purpose of comparing the linear and pattern models is that if the value of θ , the learning parameter of the linear model, is set equal to the corresponding parameter c/N (where N is the number of available patterns) in the pattern model, then the two models predict identical mean learning curves. Further, the two models agree closely for many other statistics of the data (Estes, 1959; Yellott, 1969). Thus the pattern model clearly qualified as a potential competitor to the linear model for probability learning, and, in fact, we appeared to have two alternative models of probability learning that were equally well supported by experiment.

Nonetheless, the two models are not identical, so it seemed urgent to design an experiment that could differentiate them. During the wave of probability-learning research of the 1950s, it was not apparent how a decisive test could be accomplished, partly because, while avoiding the hazards of averaging data, I had fallen into a more subtle trap. Namely, I had tacitly assumed that if curves of probability learning over trials in which responses are followed by informative feedback are not distorted materially by the use of averaged data, the same would be true if conditions were altered.

Ultimately an ingenious solution was discovered by a younger investigator, John Yellott, who was not blinded by the same tacit assumption. Yellott’s attack on the problem began with a striking mathematical relation, dubbed *mar-*

ginal constancy, that had been established in my mathematical psychology seminar at Stanford. Marginal constancy is a property of both the linear model and the pattern model, and it arises if learners have gone through a series of trials in a standard probability learning experiment and then are shifted (without their knowledge) to what is called a noncontingent success (NCS) schedule. In an NCS schedule, all responses are correct; that is, if a subject predicts event E_1 , then E_1 follows, and if the subject predicts event E_2 , then E_2 follows. Marginal constancy is a prediction of both the linear and the pattern models that the mean probability of either response, say predicting E_1 , remains constant over trials on the NCS schedule as can be seen in Figure 7.

Yellott noted a key fact, however: The two models differ in predictions about performance by individual subjects under an NCS schedule. As is illustrated in Figure 8, the linear model predicts that for any individual subject, the probability of an A_1 response will drift toward a final level of either 0 or 1. In contrast, the pattern model predicts that each individual subject will alternate indefinitely between runs of A_1 responses and runs of A_2 responses. Results of a carefully conducted test experiment decisively refuted the linear model and supported the pattern model (Yellott, 1969).

The present status of the linear model. The moral of the episode just described is that a seemingly endless piling up of correct predictions from the linear model for average data⁷ did not represent a commensurate accumulation of evidence that the model was appropriate as a representation of cognitive processes in individual learners. The ill effect of averaging in this instance was not that individual functions were distorted by averaging, but that critical features of individual behavior were obscured from the view of investigators who attended only to statistics such as means and

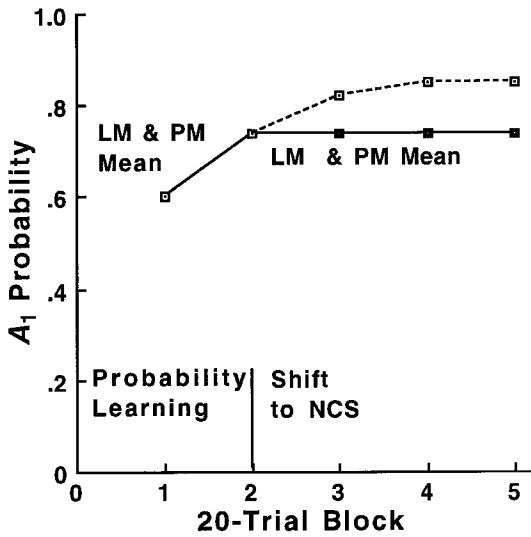


Figure 7. Learning curve for a hypothetical experiment. The first two blocks of trials are conducted under a standard probability-learning schedule with a probability of .85 for the event predicted by response A_1 but the last three blocks are under an NCS schedule in which all responses are correct. The dashed curve represents the trend over the last three blocks predicted by either the linear or the pattern model in the absence of the shift.

standard deviations that are ordinarily computed from the data of probability learning experiments. Once the blinders were removed, it was apparent that the pattern model is the more appropriate model by virtue of generating correct predictions for individual subject performance, whereas, at least under some conditions, the linear model yields correct predictions only for statistics of groups.

The linear model and probability estimation. Before concluding that the linear model has no place in a theory of probability learning, we need to avoid the trap of assuming that the type of data collected in the standard experimental paradigm fully represents all that subjects learn in the situation. In the standard paradigm, evidence as to what subjects learn about event probabilities comes only from frequencies of choice responses. However, a more direct measure merits consideration.

In a few experiments in the large literature on probability learning, subjects have had the task, not of predicting, or guessing, occurrences of alternative events, but of giving estimates, trial by trial, of the probabilities of the events (N. H. Anderson, 1969; Beach, Rose, Sayaki, Wise, & Carter, 1970; Estes, 1987; Neimark & Shuford, 1959; see also Estes, Campbell, Hatsopoulos, & Hurwitz, 1989, for a use of this task in an experiment on category learning). In each of these studies, mean learning curves for probability estimates were similar in form to those characteristics of curves for predictive behavior and exhibited close approximations to asymptotic probability matching. Thus, it was suggested that the linear model might be reinterpreted to provide an account of the learning of the mental representations that form the basis of probability estimates (N. H. Anderson, 1969; Estes, 1987). In the rein-

terpretation, the properties of the rate parameter θ in Equations 5 and 6 are unchanged, but p_n represents a response magnitude on a (0,1) scale with properties similar to those of the responses generated in psychophysical magnitude estimation (Stevens, 1957). Thus, predicted mean learning curves are identical for the two response modes—predicting events versus estimating event probabilities. The status of the linear model must be listed as “undecided” until definitive evidence is obtained concerning its adequacy for predicting performance of individual subjects in the estimation task.

A contribution from cognitive neuroscience. For a new perspective on the nature of probability learning, I will give a brief summary of a recent research finding in neural science, which at first may appear to be a digression. The report of this research, which recently came to hand under the title “The Left Hemisphere’s Role in Hypothesis Formation” (Wolford, Miller, & Gazzaniga, 2000), did not immediately suggest relevance to issues of model development. The setting for the study was an accumulation of evidence that the left frontal or prefrontal cortex is the locus of mechanisms responsible for people’s ability to test hypotheses arising from their experience. At least one of the investigators (Wolford) was also familiar with the probability-learning literature, and, in particular, with the fact that subjects in probability-learning experiments sometimes claim to be searching for causal sequences while engaged in the task.

This background led to an experiment in which split-brain patients engaged simultaneously in two probability-learning tasks, with the stimuli for one task directed to the left hemispheres of their brains and the stimuli for the other task to the right hemispheres. Learning curves for performance by the left hemispheres approached probability matching, whereas curves for the right hemispheres rose above the matching level and approached 100% predictions of the more frequent event, as is illustrated by my imaginative rendition of the gist of these findings in Figure 9.

The results for the left brains were taken to support the investigators’ expectations, on the premise that the left brains were testing hypotheses. Results for the right brains appeared to confirm the assumption that the right hemisphere is limited to a more primitive form of learning characteristic of lower animals in probabilistic situations (Hinson & Staddon, 1983). These trends were replicated with patients who had suffered unilateral damage to the right or left hemisphere.

However striking these findings may appear, it must be noted that they are not sufficient to justify the conclusion that hypothesis testing by the left brains was the necessary and sufficient condition for the tendency of their learning curves to approach probability matching. In view of the large literature on multiple, concurrent processes in human cognition (Pashler, 1993; Schweickert, 1993), we must allow for the possibility that the left brains might have been doing something besides hypothesis testing. A specific suggestion is that they might have been forming mental representations of event probabilities of the kind that have

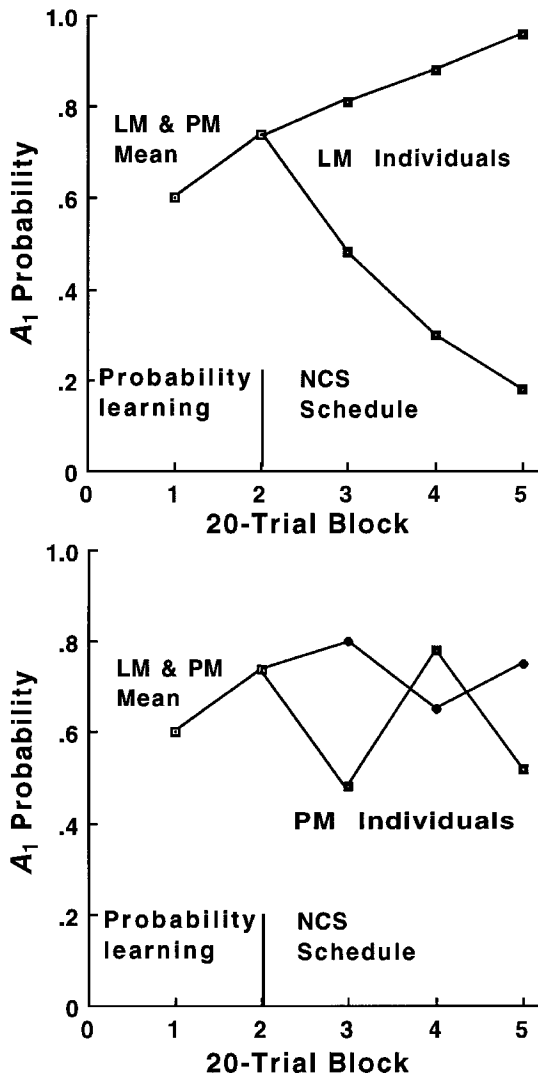


Figure 8. Learning curves for individual subjects in the experiment portrayed in Figure 7 as predicted by the linear model (upper panel) and by the pattern model (lower panel).

been revealed by experiments in which subjects give estimates, trial by trial, of to-be-predicted events. Possibly the left brains of the subjects in the Wolford et al. (2000) study were forming such representations and using them to constrain their choices of hypotheses to ones compatible with the event probabilities.

To evaluate this possibility, we evidently need a companion experiment in which split-brain and unilaterally brain-damaged patients (the same patients as in Wolford et al., 2000, if feasible) perform the probability-estimating task instead of the event-predicting task. Logically warranted conclusions about what the left brain learns in probability-learning situations wait on such a follow-up of the seminal Wolford et al. (2000) study.

Regardless of the outcome of a follow-up experiment, it is of interest to ask “What are the implications of the

findings of Wolford et al. (2000) for learning by normal individuals?” There is substantial evidence, much from behavioral research but some from neural science (e.g., evoked-potential studies), that the two hemispheres of the normal brain operate as two relatively independent systems, each of which typically contributes some of its resources to accomplishment of a cognitive task (A. Friedman & Polson, 1981). A plausible implication of the split-brain study is that in situations like probability learning, the two hemispheres of a normal individual’s brain process the incoming information in parallel, but that usually performance is governed by the dominant hemisphere and well represented by the pattern model. If, however, during a long series of trials the dominant hemisphere tires of fruitless efforts to construct hypotheses that prove uniformly incorrect, control of performance shifts to the other hemisphere, where only a more primitive mechanism of a kind shared with lower animals is available, and probability of predicting the more frequent outcome moves toward an asymptote of unity (“maximizing successes”), as has been observed in studies by Edwards (1961) and M. P. Friedman et al. (1964).

Multiple, concurrent learning processes in categorization and recognition. Taken together, the findings of Yellott (1969) and Wolford et al. (2000) point up the importance of recognizing that a single measure, such as occurrence or nonoccurrence of a predictive response, may not reflect all that is learned by a subject in an experimental task. Though these studies were all concerned with simple probability learning, the use of alternative measures of learning might be expected to be just as appropriate in research on category learning, in view of the close parallelism between standard probability-learning and category-learning experiments with respect to stimulus presentations, responses, and informative feedback.

In research on category learning, as in probability learning, appreciation of the possibility of concurrent, parallel processes did not come quickly. In one of the earliest reviews of forms of concept formation that now would be termed category learning, Hebb (1924) identified an opposition between two theoretical approaches that she denoted as “active” and “passive,” the former represented by use of rules and hypothesis testing, and the latter, by relatively automatic associative learning. Those two adjectives no longer seem appropriate, and I will substitute “hypothesis testing” and “learning.” The hypothesis-testing approach was the dominant one for more than a half century following Hebb’s review, reaching its peak of theoretical elaboration in the models of Bourne and Restle (1959), Levine (1967), and Trabasso and Bower (1968).

Beginning in the late 1970s, the long submerged learning approach was revived by the work of Rosch (1978) on category structures derived from family resemblance and the formulation of an instance-based “exemplar” model of category learning by Medin and Schaffer (1978). From that time to the present, there has ensued an almost unbroken period of dominance of the learning approach, powered mainly by the elaboration and testing of exemplar-based

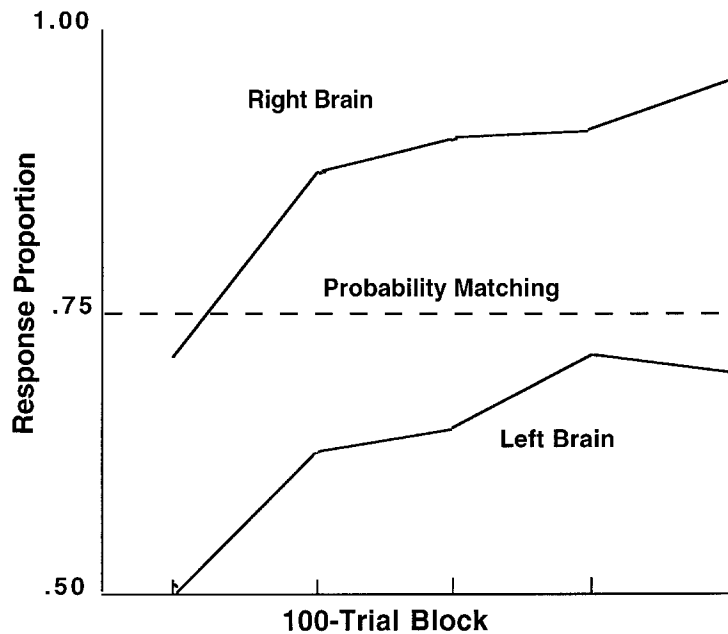


Figure 9. Learning curves of patients with split brains or with unilateral damage to the frontal cortex in a probability-learning experiment with a true probability of .75 for the predicted event. The upper and lower curves show performance under control of the right and left hemispheres of the brain, respectively.

models (Nosofsky, 1984, 1988) and closely related adaptive network models (Estes et al., 1989; Gluck & Bower, 1988; Kruschke, 1992). Questions have arisen, however, as to whether this trend will continue, owing to a development in the information-storage tradition that came as a surprise to many investigators, including myself.

Following a long succession of successful tests of exemplar-based models of category learning, Nosofsky and his associates were sufficiently impressed by subjects' reports of efforts at hypothesis testing in their experiments to formulate an alternative, rule-based model of category learning termed RULEX (Nosofsky, Clark, & Shin, 1989; Nosofsky, Palmeri, & McKinley, 1994). The RULEX model proved superior to the exemplar model at describing performance in a series of experiments on category learning, raising the possibility that the hypothesis-testing approach might have come back to represent the wave of the future.

There is need for caution about leaping to that projection, however. In research supporting both the exemplar and the RULEX models, performance has been uniformly measured in terms of binary choices between two responses (assignments of items to categories), and as in the case of the Wolford et al. (2000) study, one needs to consider the question of whether this one measure fully discloses all that the subjects are learning.

Although tracking learning solely by recording discrete choice responses has characterized nearly all model-oriented research on category learning in both the

hypothesis-testing and the learning traditions, a partial exception occurred in a study by Estes et al. (1989), in which subjects gave numerical estimates of the probability that an exemplar belonged to a category. This procedure was used for one of two groups and only on generalization tests (given without informative feedback) following each of four blocks of learning trials on which subjects of both groups made choice responses. The data for the group that gave probability estimates and the group that made binary choices on test trials affords a comparison between the alternative measures of categorization performance, illustrated in Figure 10. The widely diverging trends for the two groups leave little doubt that the two measures are tapping two distinct aspects of the learning that went on with little or no interference during the training blocks when both groups responded in the same way to the same sequences of items.⁸ Subjects giving probability estimates on the tests exhibited trends over blocks that approached close probability matching (i.e., matching of probability estimates to true probabilities), but subjects making choice responses exhibited the "overshooting" of true probabilities at the high end of the dimension and "undershooting" at the low end that has been taken to characterize rule-based behavior.

An implication of these results is that none of the experiments yielding evidence for rule-based performance have disconfirmed the assumption that subjects were simultaneously engaged in an automatic process of exemplar learning. It has to be expected that the latter process would have

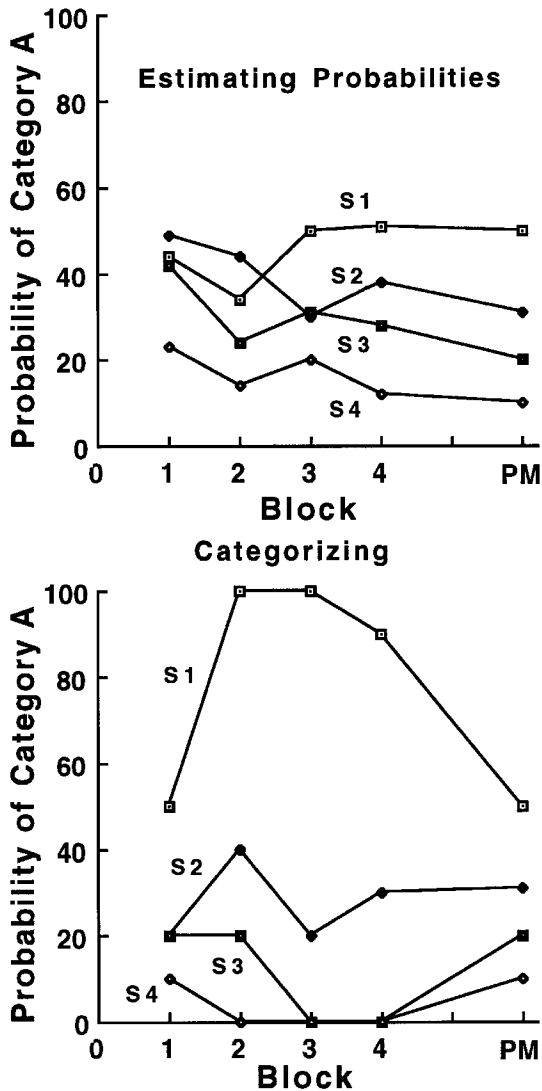


Figure 10. Mean estimates of Category A probability (upper panel) and mean proportions of Category A choices (lower panel) for the four stimulus patterns presented on test trials in the study of Estes, Campbell, Hatsopoulos, and Hurwitz (1989). True probability-matching values are plotted above the point labeled "PM" on the x-axis.

been hidden in the experiments so far conducted to test rule models, because data were collected only for choice responses, which could be based on either rules or exemplar memory.

It appears that the next wave in categorization research may be not a new hegemony of hypothesis-testing models, but a recasting of the issue of the two strains of models in terms of concurrent processes. One realization of this approach is the ATRIUM model of Erickson and Kruschke (1998). This model includes modules for rule learning and exemplar-based learning, and a graded combination of the outputs of the two modules is the basis for a predicted categorization response.

An alternative framework for handling dual concurrent processes in category learning was formulated a number of years ago by Maddox and Estes (1996). In that development, the component processes were those of an exemplar model and an adaptive network model, but the framework could as well be employed with an exemplar model and a rule-based model for components, like that shown in Figure 11.

Stimulus input on any trial is processed independently and in parallel by the rule and the exemplar modules, each producing an output in the form of a response probability (which could be the basis for either production of a category-probability estimate or a categorization response). The comparator allows the output with the more extreme value (i.e., nearer to 0 or 1) to control the response on that trial. If the response is followed by informative feedback, the state of learning in each module is updated. The critical difference from the framework of Erickson and Kruschke (1998) is that the outputs of the modules are not combined, but rather compete for control of the response-generation process. The difference should be testable by appropriate fitting of the models to individual data, perhaps using probability estimation as well as discrete responses, but the task does not look easy.

Case Study II: Models of Recognition Memory

This case study concerns models of recognition memory as they have evolved through two main phases from about 1950 to the present.⁹ In the first phase, the decision component of general signal detectability theory was imported to serve as the basis for the performance aspect of recognition. In the second phase, beginning in the 1980s, the focus shifted to exploration of mechanisms of information storage and retrieval that had emerged in the framework of information processing. The result of these two lines of development in the present scene is a number of models that have been shown to predict most of the well-established facts of recognition memory. Satisfaction with this embarrassment of riches is tempered, however, by there being apparently no prospect of constructing decisive experimental tests for relative evaluation of the models. A principal source of difficulty will be shown to be a loose coupling of the decision and memory components of the models that, at this stage, prevents any of them from being sharply disconfirmable.

Signal detection theory (SDT). In the early 1950s, a well-established mathematical theory of statistical decision was used by electrical engineers as the basis for a theory of an ideal detector—that is, a machine that would yield the best possible performance at detecting faint signals in communication networks (W. W. Peterson & Birdsall, 1953). Soon after success of this effort had been demonstrated, an engineer, Wilson P. Tanner, Jr., and a psychologist, John A. Swets, proposed that human performance in perceiving near-threshold stimuli might be described by the signal detection (henceforth, SD) model. They illustrated this proposal in terms of the task of detecting a faint flash of light against a uniformly lighted

Dual-Process Category-Learning Model

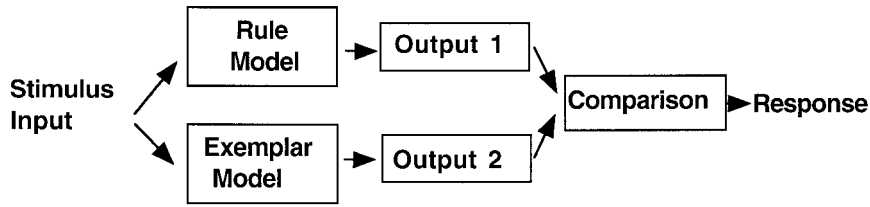


Figure 11. Theoretical framework for dual, concurrent processes in category learning (Maddox & Estes, 1996).

background. On each trial of an experiment, the experimenter either presents or omits a light flash during an observation interval and an observer (*subject* in present-day terms) attempts to report correctly whether the flash did or did not occur. The difficulty of the task arises from the fact that even a constantly illuminated background instigates random activity—"noise"—in the visual system that may be mistaken for a flash.

In their model, Tanner and Swets (1954) assumed that the parameters of an observer's visual system define a population of sensory responses ("states") that could be produced by occurrences of the flash and another population of states that could be produced by background noise during nonflash intervals of the same duration. As is illustrated in Figure 12, plots of the values of these states on an intensity scale take the forms of normal distributions with equal standard deviations but different means. The difference between the means indicates the difficulty of the task. The left-hand curve in Figure 12 represents the distribution of states produced by visual noise alone; the right-hand curve represents the distribution of states produced by the occurrence of a signal (the flash) against the noisy background (signal plus noise).

A central concept in the theory is the observer's *criterion*, indicated by the arrow in Figure 12. If the sensory state evoked in the observer on a trial falls above the criterion on the intensity scale, the observer responds "yes" to the question "Did a flash occur?" and otherwise responds "no." One speaks of a criterion as strict if it is set at a high value so that few "yes" responses will be made on noise-alone trials and as lax if it is set at a low value so that few signals will fail to produce a "yes." In many applications of SD theory, the value of the criterion is set by the investigator. The most common assumption is that the criterion is equal to 0, the mean of the noise distribution, yielding the prediction that "yes" and "no" responses on noise-alone trials will occur with equal probabilities. In other applications, the criterion is assumed to be set by the observer, possibly guided by instructions from the experimenter or by knowledge of payoffs for correct responses or penalties for errors.

Over ensuing decades, the SD model, with only technical modifications to accommodate particular applications, has become almost universally accepted as a theoretical

account of decision making in research on perceptual detection and recognition and in numerous extensions to applied domains (Swets, 1988; Swets, Dawes, & Monahan, 2000). This development may well be regarded as the most towering achievement of basic psychological research of the last half century.

The framework of SDT for transformations of data. One reason for the wide use of methods based on SDT is the possibility of obtaining useful results from their application without assuming that all of the assumptions of SDT are met. To introduce this aspect of SDT, I present a sample of performance by a subject in an experiment on word recognition. This individual had studied a list of words that included words of greatly varying degrees of familiarity, then was given a recognition test on a list containing half old and half new items with the task of responding "old" or "new" according to whether a test item was or was not judged to have come from the study list. As is shown in the upper part of Table 1, the subject judged old items of very high familiarity (VHF) to be old nearly 70% of the time. This performance might be taken to suggest fairly good performance. By itself, however, the proportion of "old" responses tells us nothing about accuracy of recognition, for the subject might merely have a general tendency (low "criterion") for choosing the re-

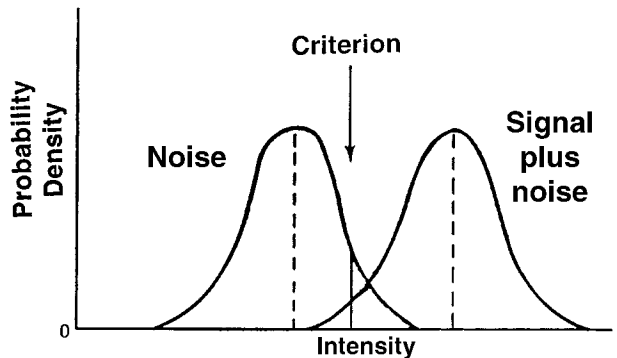


Figure 12. The decision mechanism of signal detection theory. The curves represent normal distributions with the same standard deviation but different means, one for trials on which background noise alone occurs and one for trials on which a signal is added to the noise.

Table 1
Computation of SDT Measures
for an Old–New Recognition Task

Response	Item Type	
	VHF	LF
“Old” Proportions		
Old	.69	0.75
New	.25	0.75
Standardized (z) Scores		
Old	0.49	0.67
New	–0.67	0.67
d' (Old–New)	1.16	0
$-C$ (Avg. Old, New)	–0.09	0.67

Note— d' and C are the sensitivity and criterion parameters of signal detection theory (SDT).

sponse “old” whether or not an item was recognized. A comparison is needed with responses to items of the same class that are new on the test; and the value .250 in the second row of Table 1 indicates that this subject’s tendency to respond “old” is considerably lower for new than for old items. For items of low familiarity (LF), in contrast, results in the right-hand column of the upper part of Table 1 show equal proportions of “old” responses to old and new items, suggesting no discrimination between old and new items of this category.

The necessity of comparing performance on old and new test items was recognized long before the advent of SD theory. With SD theory now available to guide the analysis, however, we can proceed to an informative quantitative summary of our results. The first step is to transform the proportions in the upper part of Table 1 to standard scores (z scores in a normal distribution with a mean of 0 and standard deviation of 1), shown in the lower part of Table 1. The value of this transformation is that it puts the observed values on a scale of measurement with advantageous properties—in particular, the property that addition and subtraction of values yield meaningful results.¹⁰ By subtracting the new from the old z scores, we obtain values of the statistic termed d' that is almost universally used as a measure of accuracy of recognition (just as it is used as a measure of accuracy in research on psychophysics and perception). The d' values obtained in this example signify moderate accuracy on VHF but only a chance level of accuracy on LF items.

By averaging the old and new z scores for each item type, we obtain a measure of the criterion for “old” responses, on the same scale as d' . In the remainder of this article, it will be convenient on occasion to replace the term *criterion*, which is closely associated with normal distribution statistics, with the more broadly applicable term *bias*. In the context of recognition performance, a bias for “old” responses (the counterpart of a low criterion) denotes a person’s tendency to choose the response “old” to a test item whether it is actually old or new. Similarly, a bias against “old” responses, the counterpart of a high criterion, denotes a tendency to avoid “old”—that is, to choose “new” responses. As can be seen in the bottom row of Table 1, the subject in

this example exhibits almost no differential bias for “old” versus “new” responses to VHF test words but a bias for “old” responses to LF words.

A critical message to carry away from this example is that the d' value computed from the SD model for a particular, previously studied test item is not a direct measure of accuracy of recognition of the item, but, rather, is an indirect measure derived from comparison of performance on that item with performance on one or more new items that were tested in the same context at nearly the same time. A similar comment applies to the criterion, or bias, measure, C .

Another point to remember is that the technique derived from SDT for transforming the data of fourfold tables is commonly used without concern for prior testing of the assumptions of the parametric, normal-distribution model of W. W. Peterson and Birdsall (1953) and Tanner and Swets (1954). Macmillan and Creelman (1991) present a thorough justification of this practice in terms of the formal properties of the measures d' and C that hold independently of the historical connection with SDT.

Extension of SDT to recognition memory. Many experimental psychologists of the mid 1950s may have noticed a resemblance between the problem facing an observer in a study of signal detection and that facing a subject in a study of recognition memory. In the former case, a test stimulus occurring in a noisy background must be distinguished from other somewhat similar stimuli that might have been generated by noise alone. In the latter case, a stimulus item that comes from a previously studied list must be distinguished from somewhat similar items that were not in the studied list. In the memory experiment, possible test items can be conceived to vary along a dimension of familiarity, just as those in the detection case vary along a dimension of intensity, and in both cases, the subject must have some criterion for dividing the populations of possible test items into those that should receive positive responses and those that should not.

To my knowledge, the first start toward a model of recognition memory based on these analogies was made by an auditory psychophysicist, James Egan (Egan, 1958). This initial foray was not followed immediately by a burst of publications in the new theoretical vein, however, and it remained for an experimental psychologist, Theodore Parks, to develop an SD-based model of recognition in detail and accompany the formulation with experimental applications (Parks, 1966). This extension of SDT evidently met a pressing need, for over ensuing decades, the extended model has become by far the most widely used conceptual tool for the interpretation and analysis of performance in research on recognition memory. The march to dominance has consisted mainly of a spreading of the range of application of the SD model, however, rather than a deepening or refining of its theoretical basis.

Components of SDT in recognition models. Beyond its heuristic value for extracting measures of discriminability and bias from fourfold tables, SDT has con-

tributed the decision mechanism that forms a component of most psychological models of recognition memory. To explicate this point, it is useful to distinguish two classes of recognition models that arise when concepts from SDT are combined with assumptions about information processing. One class I denote as process models, the other as analogical models.

A process model for recognition memory is generally conceived to be one that includes the cognitive structures that represent an individual's past encounters with objects or events and a collection of processes that enable retrieval of stored information and judgments about recognition or nonrecognition based on the retrieved information. The main purpose of developing a process model is to enable predictions of an individual's recognition performance on the basis of knowledge about the person's relevant learning history.

In contrast, an analogical model is one that is based on overall similarities, or analogies, between properties of performance on recognition tests and properties of performance in a model already constructed for some other domain. In the present context, the already available model is, of course, SDT, and one's purpose in using the analogies is to further the carryover of quantitative techniques from one domain to the other.

In the mid 1950s, there was no body of relevant theory adequate to suggest the form of a process model of recognition memory, but the way was open to use well-developed concepts of SDT as the basis for an analogical model. In a pioneering formulation of a model for performance in recognition memory tasks, Parks (1966) noted some conspicuous analogies between recognition and detection. Using the framework of Tanner and Swets (1954) as a ground plan, he defined a continuum of strength or "psychological familiarity" and assumed that the internal events generated in a subject on the study trials of a recognition experiment form distributions on this continuum. One might expect that the distributions formed during an experiment of customary duration for trials on which stimuli presented were old or new from the standpoint of the subject would look like those shown in Figure 13. The distributions actually presented by Parks, however, resembled those shown in Figure 12, essentially the noise and noise-plus-signal distri-

butions of the Tanner and Swets detection model relabeled to represent distributions on a familiarity dimension for infinite populations of items that are new or old at the time of testing. Parks assumed that the criterion for any subject is placed at the mean of the "new" distribution, the point at which "old" and "new" responses to a new test item are equally probable. In this way, the SD model of Tanner and Swets moved into psychology as an analogical model that set a template for most of the models of decision making in recognition that have appeared down to the present.

Global memory models with components from SDT. Nearly all members of the large and growing family of process models of recognition memory borrow the conceptual machinery of SDT illustrated in Figure 12 for their decision-making components. I will illustrate this tactic in terms of five of the most influential contemporary models, those of Gillund and Shiffrin (1984); Hintzman (1988); M. H. Humphreys, Bain, and Pike (1989); Murdock (1982); and Shiffrin and Steyvers (1997). All of these models share an architecture in which representations of items studied in a recognition experiment are stored as distinct traces or "images" in memory. Although the models differ with respect to assumptions about encoding of items and details of the retrieval process, all assume that activation of the stored traces in parallel with presentation of a probe item on a recognition test provides the basis for a decision about recognition (hence the familiar label, "global models"). The level of activation defines a value for the item on a "familiarity" dimension. In each of the models, the subject sets a criterion value, C , on this dimension and responds "old" to the probe if its value is greater than C , "new" if its value is below C . The assumed flow of events on a trial is illustrated in Figure 14.

In the model of Gillund and Shiffrin (1984), items presented during study are represented in memory in the form of images, defined for computational purposes as sets of features. On a recognition test, memory is probed by presentation of a test item, which raises the level of activation of all of the stored images. The subject responds "old" (the item was in the study list) if the activation level exceeds a criterion level, C , and otherwise responds "new" (the item was not in the study list). The value of C is assumed to be chosen by the subject, but how the subject

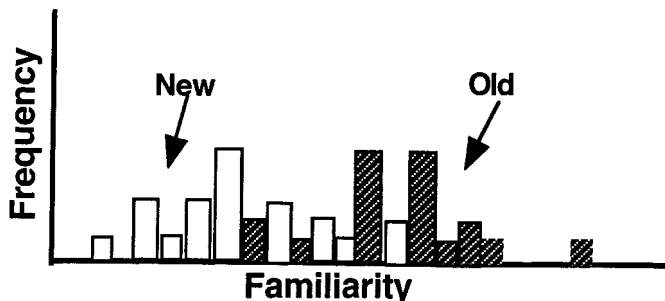


Figure 13. Frequency distribution on a familiarity dimension for memory traces of new items (white bars) and old items (cross-hatched bars) in a hypothetical old–new recognition memory experiment.

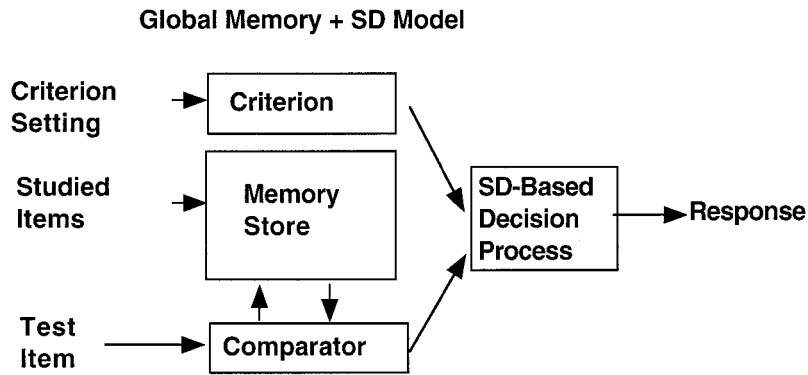


Figure 14. Schema of a global model of recognition memory with a decision mechanism drawn from signal detection theory. Presentation of a test item generates a value of memory activation that is compared with a criterion set by the subject in order to produce a response.

learns where to place C on the familiarity dimension is left an open question.

The MINERVA model of Hintzman (1988) is generally similar in structure. The stored representations (traces) of studied items are defined as vectors of feature loadings with values of 1, 0, or -1 for each feature, but as in the model of Gillund and Shiffrin (1984), feature values are not specifically related to stimulus properties. On a recognition test, presentation of a test item activates all of the traces in parallel, generating what Hintzman terms an “echo,” whose value on an echo-intensity scale is a function of the overall similarity of the test item to the traces. A criterion is defined on this scale, and an “old” or “new” response is generated on a test according to whether the echo falls above or below the criterion. Hintzman notes that an important shortcoming of the model is its lack of an account of how subjects set criteria.

In the REM model of Shiffrin and Steyvers (1997), memory of perceived items consists of stored images which take the form of error-prone feature vectors, with positive integers as entries (zeros for missing features). As in both of the other models cited, the features are abstract and not defined in terms of stimulus properties. On a test trial of a recognition experiment, the probe item is compared with all of the images in memory in parallel, and for each comparison, the likelihood is calculated that the observed degree of match would occur if the image did or did not represent the same item as the probe. Based on the results of these computations, an application of Bayes’s rule gives the probability that the probe is old—that is, that it corresponds to at least one of the stored images. The subject sets a criterion at the value .5 on the probability scale and makes an “old” response if the computed probability exceeds .5, otherwise a “new” response. The criterion value is chosen on normative grounds—that is, because it will lead to optimal performance in the long run. How the subject develops the capability of doing mental computations that, in effect, lead to correct application of

Bayes’s rule and optimal criterion setting remains an open question for future research.

Whereas in each of the models just cited, items are stored in memory as distinct images, Murdock (1982) assumes that items are stored in a distributed memory where they lose their individual identities. Recognition depends on activation of memory by a test item; the result plus unrelated noise enters a decision system where it is compared with criteria set by the subject.

The MATRIX model of M. H. Humphreys et al. (1989) also employs a distributed memory. The model, intended to encompass a wide variety of phenomena of memory, is much too extensive to be summarized here, but as applied to recognition, reduces to a version of the model of Murdock (1982).

It is natural to wonder whether appropriate experimental tests could not reduce this proliferation of models by showing that some one of them is superior. Prospects for that direct approach do not seem promising, however. One difficulty is that any two of the models differ in more than one assumption about structures or processes, so that a comparison is not informative about the components of the models. A second difficulty is that evidence for all of these models comes mainly from fits of models to averaged data, and there is substantial evidence (Ashby, Maddox, & Lee, 1994; Maddox, 1999) that the kinds of distortion by averaging that I have cited in connection with simple learning curves must be a similar hazard for interpretation of the theoretical functions that occur in global models of memory. A third difficulty with the direct approach bears on an aspect of all of the global models that needs more detailed discussion.

In each of the global models cited, like many others that could be added to the list, presentation of a probe item on a recognition test generates a mental representation that may be viewed as a value on a familiarity scale, and an account of the way in which this outcome is generated qualifies as a process model of the information storage and re-

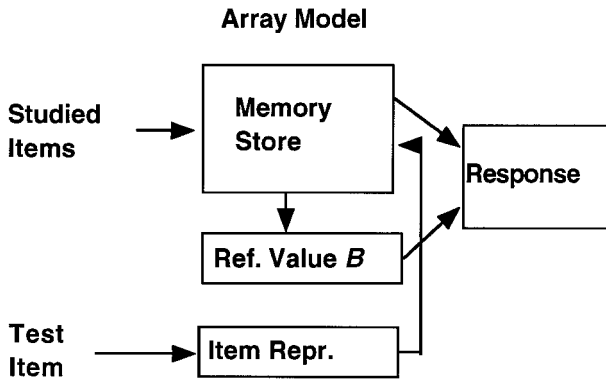


Figure 15. Schema of the array model of recognition memory. Presentation of a test item generates a value of similarity to memory that is compared with a reference level *B* in order to produce a response. The value of *B* is a product of the learning process of the model as it has operated prior to the test.

trieval components of recognition. However, a subject's decision as to whether the computed familiarity value should lead to an "old" or a "new" response is predicted by the SD decision model, appended to the incomplete process model.

For as long as I have followed the development of these and other global models of recognition memory, I have been struck by the sharp dichotomy between two modes of thought. Generation of the familiarity value, or the equivalent, of a test item is assumed to be the outcome of a causal chain of processes leading from a subject's perception of the item to the final step in a sequence of computations. Generation of a response, however, depends on a mental act of criterion setting by the subject, unconnected to those processes. In applications of the various models, subjects are sometimes assumed to choose uniform criterion settings in accord with considerations of rationality, sometimes to vary their criterion settings flexibly from one experimental condition to another in the ways needed to account for data patterns.¹¹ It is not always clearly realized that experimental tests of memory models often are tests of the structural and processing assumptions of the model combined with assumptions about criterion setting chosen to fit a particular situation. In terms of the rules of logical inference, successful predictions of data in any one of these tests yields evidence that the model is sufficient, but not that it is necessary to account for the data.

The array model. I am led to raise the question of whether it would be possible to construct a model that would share the capability of the global models for describing and predicting recognition data but would not mix causal chains and acts of free will in the manner that has become customary. To provide one indication that the goal might be feasible, I will sketch still another global memory model, termed the *array model*, that is under development by W. Todd Maddox and myself (Estes, 1994; Estes & Maddox, 1995; Maddox & Estes, 1997). The architecture and

assumptions about information storage and retrieval are generally similar to those of the global memory models described above, and, as is illustrated in Figure 15, the flow of events during a trial is the same except for the absence of a criterion-setting module. Traces of items presented for study in a recognition experiment are stored in an "old" memory array, encoded in terms of their perceptual or semantic features. On a test trial, similarities of the test item to all of the stored traces are computed, and comparison of the sum of these similarities (termed "similarity to memory") with a reference level provides the basis for recognition.

In general, the features of the model are not treated simply as abstract entities, but are set into correspondence with perceptual or other properties of the stimuli used as items whenever there is some accepted way of doing so. The purposes of this article will be served, however, by limiting consideration of the array model to a very special case in which stimuli are treated as patterns and a single parameter, *s*, with a value between 0 and 1, is used as a measure of average similarity between any two nonidentical items.

Application of this special case of the model to recognition proceeds as follows. In response to the task instructions, the learner sets up a memory array, and after *N* items have been presented for study, their *N* representations are stored in this array. If one of these items is presented on an old–new recognition test, its total similarity to the memory array is computed. The result is termed Sim_o , where the subscript identifies an old item, and in the example is equal to $(1 + s + s \dots) = 1 + s(N - 1)$. If a new (unstudied) item were tested, its similarity to memory, termed Sim_n , would be equal to sN .

Probabilities of "old" responses to old or new test items are obtained by entering Sim_o or Sim_n , respectively, into a general formula derived from instance-based models of categorization (Estes, 1986; Nosofsky, 1984, 1988). The expressions obtained for old and new tests are, respectively,

$$P_o(\text{Old}) = Sim_o / (Sim_o + B), \tag{7}$$

and

$$P_n(\text{Old}) = Sim_n / (Sim_n + B), \tag{8}$$

where *B* is a reference value with which the similarity of a test item to the memory array is compared as the basis for generating a recognition judgment.

In all versions of the model, the reference value *B* is assumed to be the same for tests of old and new items because the learner cannot know prior to presentation of an item whether it will be old or new. Functionally, the parameter *B* is an index of response bias, for as its value increases, probabilities of "old" responses to both old and new test items decrease, regardless of whether the items are old or new at the time of testing. However, the role of *B* differs from that of criterion parameters in other models, for in the array model, the value of *B* is not selected by a mental act of criterion setting on the part of the learner (or the assumption of such an act by the investigator).

To bring out the way in which the value of B is constrained by the architecture of the model, we rearrange the terms of Equation 1 as follows:

$$\text{Sim}_o + B = \text{Sim}_o/P_o(\text{Old})$$

and therefore,

$$\begin{aligned} B &= [\text{Sim}_o/P_o(\text{Old})] - \text{Sim}_o \\ &= \text{Sim}_o[\{1/P_o(\text{Old})\} - 1] \\ &= \text{Sim}_o[1 - P_o(\text{Old})]/P_o(\text{Old}). \end{aligned} \quad (9)$$

An identical series of manipulations of Equation 2 yields

$$B = \text{Sim}_n[1 - P_n(\text{Old})]/P_n(\text{Old}). \quad (10)$$

Substituting this result in Equation 1 enables it to be rewritten in the form

$$P_o(\text{Old}) = \text{Sim}_o / [\text{Sim}_o + \text{Sim}_n \{ (1 - P_n(\text{Old})) / P_n(\text{Old}) \}]. \quad (11)$$

The term $[1 - P_n(\text{Old})]/P_n(\text{Old})$ will be recognized as an expression of the mathematical odds (from the standpoint of a learner, in the framework of the model) that the test item is new. Thus, in effect, the learner's probability of calling an old test item "old" results from a comparison of the item's similarity to the memory array with the similarity that would have obtained for a new item on the same trial, weighted by the prior expectation that the item would be new. The flow of events in the array model, illustrated in Figure 15, will be seen to be simpler than that of the other global models (Figure 14) in that the array model includes no supernumerary in the form of a criterion setter.

It remains to be shown how the reference value B is related to the criterion parameter C of SD theory. For this purpose, I will use a sample of data from an experiment on word recognition in a study by Estes and Maddox (in press).¹² The stimuli were words drawn from four levels of normative word frequency in sources such as Kučera and Francis (1970)¹³ together with nonwords matched with the words for length and normative letter frequencies. Samples of nonwords and words of each of the frequency levels (henceforth denoted NW for nonwords and VLF, LF, HF, and VHF for very low, low, high, and very high frequencies) were presented for study, and subjects were then tested for old–new recognition on a mixture of half old (studied) items and half new (unstudied) items. The array model was fitted to the hit and false alarm data obtained at each normative-frequency category for each individual subject. The results, in terms of mean observed and predicted proportions of hits and false alarms, are shown in Figure 16.

The purpose of model fitting in this instance was not, however, to compare the results with fits of other models, but to enable us to see whether the reference parameter B of the array model measures response bias in the same way as does the criterion parameter C of SDT. The first step was to discover how values of B , estimated from the data, correlate with the criterion parameter C of SD theory, estimated from the same data. For the first step in this comparison, esti-

mates of B from the array model fits and estimates of C computed in the standard manner from observed hit and false alarm proportions are shown in Figure 17.¹⁴

For a second step in the comparison, Figure 18 presents the correlations (Pearson r s) between estimates of B and C for the 37 individual subjects in our sample. Except for three "outliers" whose low correlations we cannot account for, there is quite close agreement between the two measures at the level of individual subject data. More cases need to be studied, of course, but the parallelism of the trends for B and C in Figure 17, together with the correlations, is encouraging for the view that a measure of response bias generated by an undiluted memory-based model of recognition may have properties similar to those of a standard measure generated by application of SDT to the same data.

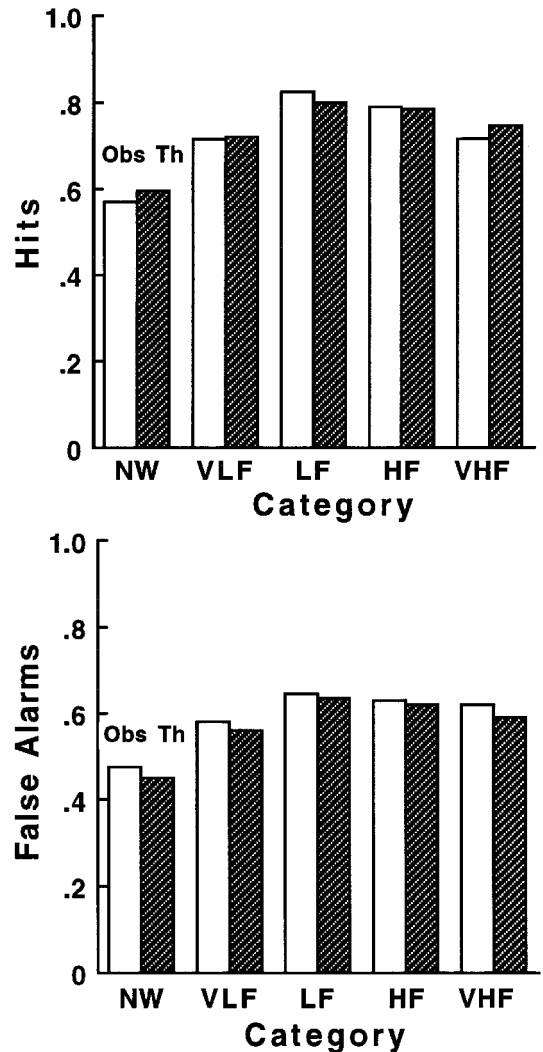


Figure 16. Observed and theoretical proportions of hits and false alarms from an experiment on recognition memory for nonwords (NW) and words sampled from four normative frequency categories (very low, low, high, and very high). Theoretical values were computed from the array model.

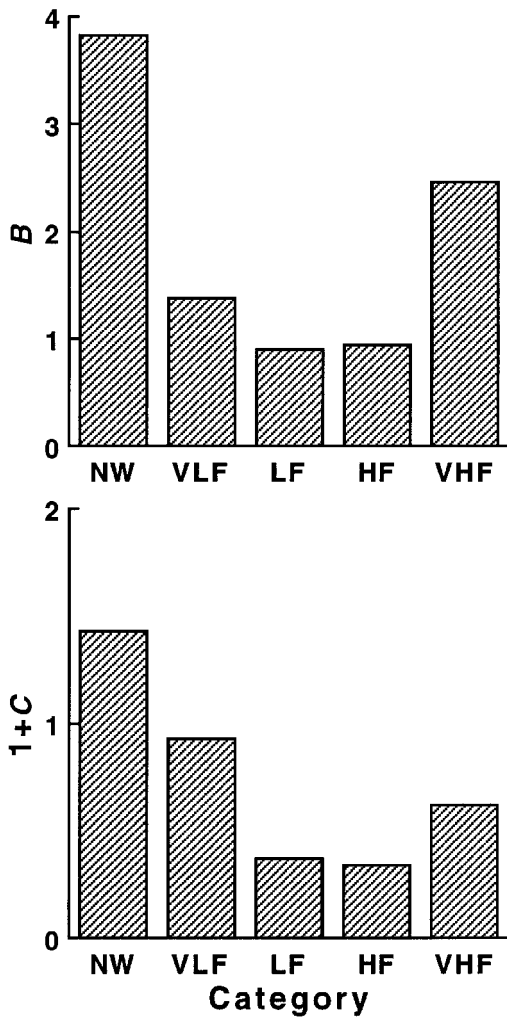


Figure 17. Trends over the nonword and word categories of Figure 16 for the reference parameter B of the array model (upper panel) and $1 + C$, where C is the criterion parameter of signal detection theory.

I do not suggest that the reference parameter B of the array model will prove to be as widely applicable as the criterion parameter C of SDT. However, I think I can safely propose that the array model, with the reference value determined by the information-processing mechanisms of the model, is more sharply disconfirmable by experiment than the other global models are.

The primary objective of this last exercise has been to illustrate a special value in having parameters defined in terms of underlying structures or processes. With this orientation, estimates of parameter values can be more informative than raw statistics of data with respect to theoretical issues. Showing that a model describes data is not an end in itself, but, rather, a preliminary to use of the model as an aid to extracting generalizable information from the empirical measures that constitute the immediate products of research.

Fortunately, the likelihood, perhaps certainty, of ultimate disconfirmation of any one version of the array model does not hinder its usefulness for our main purpose of testing particular model components. I will illustrate this point with an effort by Maddox and myself to evaluate assumptions about the learning process that occurs in study phases of recognition experiments. It is assumed in all of the global models that on each item presentation during a study phase, a representation of the item is stored in memory with some fixed probability, usually taken to be unity. In research on category learning, the adequacy of this assumption has been approached by comparing models that differ by inclusion or exclusion of a mechanism termed the “delta rule” (Gluck & Bower, 1988; Stone, 1986), which implies that efficiency or probability of storage increases with the novelty or “surprise value” of an item for a learner.

For a first exploration in research on recognition memory, Maddox and I assumed that probability of storage of an item is maximal on its first occurrence in a context and declines over repetitions according to the function α^n ($n = 0, 1, 2, \dots$), where n denotes the number of previous occurrences of the item. In applications of the array model to several data sets, we have uniformly found a substantial advantage in goodness of fit for the model version including this assumption over an otherwise identical version that did not.

Following the approach illustrated in this section, we have shifted from the traditional goal of model testing, which here would be rejecting the array model’s competitors, to the goal of gaining evidence about the generality of particular assumptions that might be advantageously included in any member of the family of alternative global models for recognition memory, and perhaps also in exemplar-based models of category learning.

Afterthoughts

Model evaluation: Reprise. I trust that the preceding sections of this article have made the point that describing data by means of a model is not an end in itself, but, rather, a preliminary to use of the model as an aid to abstracting generalizable information from the immediate products of research. Still, the preliminary is important, and it is timely to ask what can be said about the state of the art of model evaluation some 50-odd years after the construction and use of computational models of learning, memory, and decision emerged as a significant component of research. I will hazard some comments on several aspects of this question, based on impressions I have formed as a participant in this development and as a reader of much of the literature produced.

Logic of model testing. Familiarity with the relevant logical rules of inference appears to be fairly widespread in the cognitive research community, but day-to-day application of them does not. Finding that one’s model yields predictions of a set of data that are correct by some informal criterion seems often to generate a subjective feeling of confidence in the appropriateness (or long-term viability) of the model that defies reason. Too rarely do we see the au-

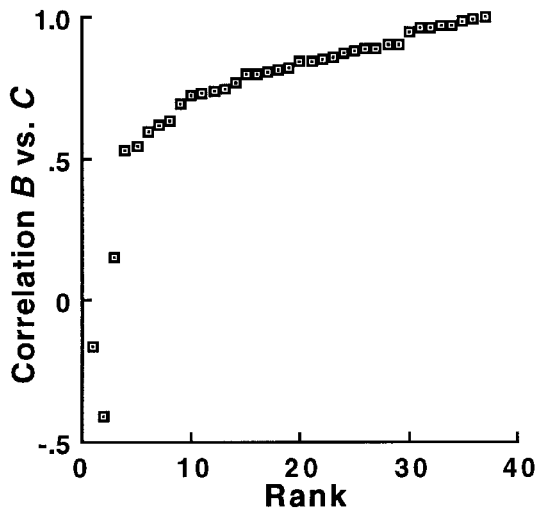


Figure 18. Correlations between the reference parameter B of the array model and the criterion parameter C of signal detection theory for individual subjects, ranked by magnitude.

thor of a report of a new model go beyond demonstrations of its success at predicting available data to investigate the consequences of altering or deleting specific assumptions about cognitive structures or processes (i.e., to consider necessity as well as sufficiency of the assumptions).

The signal-plus-noise problem. Behavioral data always include substantial error components, and the best-fitting model in a situation may simply be the one that best capitalizes on a favorable sample of error. There have been major advances in understanding the nuances of this problem for model evaluation, but a full grasp of these advances is so far limited to a few experts. The section on Coping With Error in this article (pp. 5–7) is intended to help investigators of memory and decision cope with the problem at a practical level.

General versus limited models. Cumulative progress depends on going beyond one-shot efforts and addressing all of the properties desired of a model. A successful effort typically starts with outlining of a framework, usually in the form of a general model, that presents assumptions intended to hold throughout an empirical domain. The general model is too complex to be testable as a unit, but it sets the stage for the derivation of more limited submodels amenable to quantitative testing in limited portions of the domain. The result is a family of interrelated, limited models, sometimes deserving the appellation “laws,” that preserve in distilled form the generalizable knowledge generated by the research program. Examples of this strategy in action are the stimulus-sampling framework from which stemmed the probability-learning models discussed in Case Study I, the Search of Associative Memory model of Shiffrin and associates (Gillund & Shiffrin, 1984; Raaijmakers & Shiffrin, 1981; Shiffrin, Ratcliff, & Clark, 1990), the ACT* model of J. R. Anderson (1983), and the parallel-processing framework of Rumelhart, Hinton, and McClelland (1986).

Logic and creativity. One often encounters the view that logic and creativity are antithetical: Significant discoveries and great theories are said to be due to inspiration and intuition, not to mechanical processes of model building and testing. On this issue, I have two comments. First, the evaluation of models is by no means a mechanical process of applying logical algorithms. Though constrained by logic, preferences among alternative models are often based on educated guesses as to which can best be extended in a natural way to related phenomena beyond the range of the original tests or which will, in applications, yield more new insights or other informative interpretations of data. Finally, as I have tried to illustrate in my Case Studies I and II, the formulation and testing of appropriate models constitutes a main route to an understanding of phenomena that goes deeper than compilations of empirical facts.

Models in cognitive and neural science. People for whom the technicalities of model construction and testing are uncongenial may be resting comfortably with the idea that advances in neural science will soon make models unnecessary in cognitive science. For disillusionment of this view, I suggest a look at a recent special issue of the journal *Nature Neuroscience* (November 2000, Volume 3 Supplement) devoted to computational approaches to brain function. In a Foreword, one learns that four institutes of the U. S. National Institutes of Health, sponsor of the special issue, have formed or are currently forming programs in theoretical and computational neuroscience, every one of which will be focusing on interactions between models and experimental research in problem areas that span neural, behavioral, and cognitive science. Developments in usage of models reviewed in the special issue range from research on synaptic plasticity and cerebellar function to research on working memory and selective attention. I hope that progress in these new efforts may be furthered if investigators on both sides of the boundary between brain and behavior are sensitized to the logical and statistical traps that can impede fruitful interactions between models and data at any level of theory construction.

REFERENCES

- ANDERSON, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.
- ANDERSON, N. H. (1969). Application of a model for numerical responses to a probability learning situation. *Journal of Experimental Psychology*, **80**, 19-27.
- ANDERSON, R. B., & TWENEY, R. D. (1997). Artfactual power curves in forgetting. *Memory & Cognition*, **25**, 724-730.
- ASHBY, F. G., & ALPHONSO-REESE, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, **39**, 216-233.
- ASHBY, F. G., MADDUX, W. T., & LEE, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science*, **5**, 144-151.
- ATKINSON, R. C., CARTERETTE, E. C., & KINCHLA, R. C. (1962). Sequential phenomena in psychophysical judgments: A theoretical analysis. In E. D. Neimark & W. K. Estes (Eds.), *Stimulus sampling theory* (pp. 601-665). San Francisco: Holden Day.
- BAKAN, D. (1954). A generalization of Sidman's results on group and individual functions, and a criterion. *Psychological Bulletin*, **54**, 63-64.

- BEACH, L. R., ROSE, R. M., SAYAKI, Y., WISE, J. A., & CARTER, W. B. (1970). Probability learning: Response proportions and verbal estimates. *Journal of Experimental Psychology*, **86**, 165-170.
- BOURNE, L. E., & RESTLE, F. (1959). A mathematical theory of concept identification. *Psychological Review*, **66**, 278-296.
- BRUNSWIK, E. (1939). Probability as a determiner of rat behavior. *Journal of Experimental Psychology*, **25**, 175-197.
- BURKE, C. J. (1959). Applications of a linear model to two-person interactions. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory* (pp. 180-203). Stanford, CA: Stanford University Press.
- BUSH, R. R. (1963). Estimation and evaluation. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 430-469). New York: Wiley.
- BUSH, R. R., & MOSTELLER, F. (1955). *Stochastic models for learning*. New York: Wiley.
- DETABEL, M. H. (1955). A test of a model for multiple-choice behavior. *Journal of Experimental Psychology*, **49**, 97-104.
- DORFMAN, D. D. (1969). Probability matching in signal detection. *Psychonomic Science*, **17**, 103.
- EBBINGHAUS, H. (1885). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie* [On memory: Investigations in experimental psychology]. Leipzig: Dunker & Humboldt.
- EBBINGHAUS, H. (1964). *Memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenius, Trans.). New York: Dover. (Original work published 1885)
- EDWARDS, W. (1961). Probability learning in 1,000 trials. *Journal of Experimental Psychology*, **62**, 385-394.
- EGAN, J. P. (1958). *Recognition memory and the operating characteristic* (Tech. Rep. No. AFCRC TN-58-51, AD-152650). Bloomington: Indiana University, Hearing and Communication Laboratory.
- ERICKSON, M. A., & KRUSCHKE, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, **127**, 107-140.
- ESTES, W. K. (1950). Toward a statistical theory of learning. *Psychological Review*, **57**, 94-104.
- ESTES, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, **53**, 134-140.
- ESTES, W. K. (1957). Theory of learning with constant, variable, or contingent probabilities of reinforcement. *Psychometrika*, **22**, 113-132.
- ESTES, W. K. (1959). Component and pattern models with Markovian interpretations. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory* (pp. 9-52). Stanford, CA: Stanford University Press.
- ESTES, W. K. (1960). 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. (1986). Storage and retrieval processes in category learning. *Journal of Experimental Psychology: General*, **115**, 155-174.
- ESTES, W. K. (1987). Application of a cognitive distance model to learning in a simulated travel task. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **13**, 380-386.
- ESTES, W. K. (1994). *Classification and cognition*. Oxford: Oxford University Press.
- ESTES, W. K., CAMPBELL, J. A., HATSOPOULIS, N., & HURWITZ, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 556-571.
- ESTES, W. K., & JOHNS, M. (1958). Probability learning with ambiguity in the reinforcing stimulus. *American Journal of Psychology*, **71**, 219-228.
- ESTES, W. K., & MADDOX, W. T. (1995). Interactions of similarity, base rate, and feedback in recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 1075-1095.
- ESTES, W. K., & MADDOX, W. T. (in press). On the processes underlying stimulus-familiarity effects in recognition of words and nonwords. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**.
- ESTES, W. K., & STRAUGHAN, J. H. (1956). Analysis of a verbal conditioning experiment in terms of statistical learning theory. *Journal of Experimental Psychology*, **47**, 225-234.
- ESTES, W. K., & SUPPES, P. (1959). Foundations of linear models. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory* (pp. 137-179). Stanford, CA: Stanford University Press.
- FRIEDMAN, A., & POLSON, M. C. (1981). Hemispheres as independent resource systems: Limited-capacity processing and cerebral specialization. *Journal of Experimental Psychology: Human Perception & Performance*, **7**, 1031-1058.
- FRIEDMAN, M. P., BURKE, C. J., COLE, M., KELLER, L., MILLWARD, R. B., & ESTES, W. K. (1964). In R. C. Atkinson (Ed.), *Studies in mathematical psychology* (pp. 250-316). Stanford, CA: Stanford University Press.
- GALLISTEL, C. R. (1990). *The organization of learning*. Cambridge, MA: MIT Press.
- GILLUND, G., & SHIFFRIN, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, **91**, 1-67.
- GLUCK, M. A., & BOWER, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, **117**, 227-247.
- GRANT, D. A., HAKE, H. W., & HORNSETH, J. P. (1951). Acquisition and extinction of a verbal conditioned response with differing percentages of reinforcement. *Journal of Experimental Psychology*, **42**, 1-5.
- GRUNWALD, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, **44**, 133-152.
- HEATHCOTE, A., BROWN, S., & MEWHORT, D. J. K. (2000). The power law revealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, **7**, 185-207.
- HEIDBREDE, E. (1924). An experimental study of thinking. *Archives of Psychology*, **11** (No. 73).
- HILGARD, E. R. (1951). Methods and procedures in the study of learning. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 517-567). New York: Wiley.
- HINSON, J. M., & STADDON, J. E. R. (1983). Matching, maximizing, and hill-climbing. *Journal of the Experimental Analysis of Behavior*, **40**, 321-331.
- HINTZMAN, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory. *Psychological Review*, **96**, 528-551.
- HIRSHMAN, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 302-313.
- HUMPHREYS, L. G. (1939). Acquisition and extinction of verbal expectations in a situation analogous to conditioning. *Journal of Experimental Psychology*, **25**, 294-301.
- HUMPHREYS, M. H., BAIN, J. D., & PIKE, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review*, **96**, 208-233.
- KINCHLA, R. A. (1962). *Learned factors in psychophysical discrimination*. Unpublished doctoral dissertation, University of California, Los Angeles.
- KOFFKA, K. (1924). *The growth of the mind* (R. M. Ogden, Trans.). London: Kegan Paul, Trench, Trubner.
- KRUSCHKE, J. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, **99**, 22-44.
- KUCERA, H., & FRANCIS, W. N. (1970). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
- LEVINE, M. (1967). The size of the hypothesis set during discrimination learning. *Psychological Review*, **74**, 428-430.
- MACMILLAN, N. A., & CREELMAN, C. D. (1991). *Detection theory: A user's guide*. Cambridge: Cambridge University Press.
- MADDOX, W. T. (1999). On the dangers of averaging across observers when comparing decision bound and generalized context models of categorization. *Perception & Psychophysics*, **61**, 354-374.
- MADDOX, W. T., & ESTES, W. K. (1996, August). *A dual-process architecture for models of category learning*. Paper given at the 29th Annual Meeting of the Society for Mathematical Psychology, Chapel Hill, NC.
- MADDOX, W. T., & ESTES, W. K. (1997). Direct and indirect stimulus-frequency effects in recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 539-559.
- MC GEOCH, J. A. (1942). *The psychology of human learning*. New York: Longmans, Green.

- MEDIN, D. L., & SCHAFFER, M. M. (1978). Context theory of classification learning. *Psychological Review*, **85**, 207-238.
- MURDOCK, B. B., JR. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, **89**, 609-626.
- MYUNG, J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, **44**, 190-204.
- NEIMARK, E. D., & SHUFORD, E. H. (1959). Comparison of predictions and estimates in a probability learning situation. *Journal of Experimental Psychology*, **57**, 294-298.
- NEWELL, A., & ROSENBLOOM, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-55). Hillsdale, NJ: Erlbaum.
- NOSOFSKY, R. N. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 104-114.
- NOSOFSKY, R. N. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 700-708.
- NOSOFSKY, R. N. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*, **43**, 211-233.
- NOSOFSKY, R. N., CLARK, S. E., & SHIN, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 282-304.
- NOSOFSKY, R. N., PALMERI, T. J., & MCKINLEY, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, **101**, 53-79.
- PARKS, T. E. (1966). Signal detectability theory of recognition-memory performance. *Psychological Review*, **73**, 44-58.
- PASHLER, H. (1993). Dual-task interference and elementary mental mechanisms. In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 245-264). Cambridge, MA: MIT Press.
- PETERSON, C., & BEACH, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, **68**, 29-46.
- PETERSON, W. W., & BIRDSALL, T. G. (1953). *The theory of signal detectability*. (Tech. Rep. No. 13). University of Michigan: Electronic Defense Group.
- PITZ, G. F. (1980). The very guide of life: The use of probabilistic information for making decisions. In T. W. Wallsten (Ed.), *Cognitive processes in choice and decision behavior* (pp. 77-94). Hillsdale, NJ: Erlbaum.
- RAAIJMAKERS, J. G. W., & SHIFFRIN, R. M. (1981). Search of associative memory. *Psychological Review*, **88**, 93-134.
- ROSCH, E. (1978). Principles of categorization. In E. Rosch & E. E. Lloyd (Eds.), *Cognition and categorization* (pp. 27-48). Hillsdale, NJ: Erlbaum.
- RUMELHART, D. E., HINTON, G. E., & MCCLELLAND, J. L. (1986). A general framework for parallel processing. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations* (pp. 77-109). Cambridge, MA: MIT Press.
- SCHWEICKERT, R. (1993). Information, time, and the structure of mental events. In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 535-566). Cambridge, MA: MIT Press.
- SHEPARD, R. N. (1958). Stimulus and response generalization: Deduction of the generalization gradient from a trace model. *Psychological Review*, **65**, 242-256.
- SHEPARD, R. N. (1974). Representation of structure in similarity data. *Psychometrika*, **39**, 373-421.
- SHEPARD, R. N., ROMNEY, A. K., & NERLOVE, S. (1972). *Multidimensional scaling theory: Theory and applications in the behavioral sciences*. New York: Academic Press.
- SHIFFRIN, R. M., RATCLIFF, R., & CLARK, S. E. (1990). The list-strength effect: Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 179-95.
- SHIFFRIN, R. M., & STEYVERS, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, **4**, 145-166.
- SIDMAN, M. (1952). A note on functional relations obtained from group data. *Psychological Bulletin*, **49**, 263-269.
- SLOVIC, P., LICHTENSTEIN, S., & FISCHHOFF, B. (1988). Decision making. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey, & R. D. Luce (Eds.), *Stevens' Handbook of experimental psychology: Vol. 2. Learning and cognition* (2nd ed., pp. 673-738). New York: Wiley.
- STEVENS, S. S. (1957). On the psychophysical law. *Psychological Review*, **64**, 153-181.
- STONE, G. O. (1986). An analysis of the delta rule and the learning of statistical associations. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations* (pp. 444-459). Cambridge, MA: MIT Press.
- STRETCH, V., & WIXTED, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 1397-1410.
- SUPPES, P. (1957). *Introduction to logic*. Princeton, NJ: Van Nostrand.
- SUPPES, P., & ATKINSON, R. C. (1960). *Markov learning models for multiperson interactions*. Stanford, CA: Stanford University Press.
- SWETS, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285-1293.
- SWETS, J. A., DAWES, R. M., & MONAHAN, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, **1**, 1-26.
- TANNER, W. P., JR., & SWETS, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, **61**, 401-409.
- TRABASSO, T., & BOWER, G. H. (1968). *Attention in learning: Theory and research*. New York: Wiley.
- WICKENS, T. D. (1982). *Models for behavior*. San Francisco: Freeman.
- WIXTED, J. T. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 681-690.
- WIXTED, J. T., & EBBESEN, E. B. (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory & Cognition*, **25**, 731-739.
- WOLFORD, G., MILLER, M. B., & GAZZANIGA, M. (2000). The left hemisphere's role in hypothesis formation. *Journal of Neuroscience*, **20** (RC64), 1-4.
- YELLOTT, J. I., JR. (1969). Probability learning with noncontingent success. *Journal of Mathematical Psychology*, **6**, 541-575.

NOTES

1. Illustrations of some of these conditions are given in the Appendix.
2. In this context, power and exponential curves for retention typically take the forms $P = at^{-b}$ and $P = ae^{-bt}$, respectively, where P is a performance measure, t is duration of a retention interval, and a and b are constants to be estimated from data. Curves for reaction time during practice are $RT = a + bn^{-c}$ and $RT = a + be^{-cn}$, respectively, where RT denotes reaction time, n denotes trial number, and a , b , and c are constants.
3. The three exponential functions differ only in the values (.1, .2, and .3) of the rate constant, c .
4. Both articles are included in a special issue of the *Journal of Mathematical Psychology* (March, 2000), devoted to model selection. The article by Myung (2000) is a good tutorial for general readers.
5. This estimate is based only on my own research experience but seems compatible with the analyses reported by Myung (2000).
6. In the terminology I now prefer, the expression "S is associated with E_i " translates to "representations of S and E_i are stored together as a single entry in a memory array."
7. The reference is to numerous variations on the basic probability-learning paradigm in which the linear model predicted learning curves with notable accuracy (Detambel, 1955; Estes, 1957; Estes & Johns, 1958; Kinchla, 1962; Neimark & Shuford, 1959).
8. Data are from Estes et al. (1989, Experiment 2, group having no informative feedback on test trials).

- 9. Models limited strictly to short-term recognition are excluded.
- 10. This transformation is good practice even if SDT is not involved but a parametric statistical analysis like ANOVA is to be used.
- 11. Advances in the investigation of regularities in criterion setting (Hirshman, 1995; Stretch & Wixted, 1998; Wixted, 1992) are largely being accomplished outside of the frameworks of formal models other than SDT.

- 12. This sample constitutes the *lexical* group in Experiment 1 of Estes and Maddox (in press).
- 13. Normative frequencies were, for our very low level, less than 1 in 6 million; low level, 1 per million; high, 2–39 per million; very high, 41–2,714 per million.
- 14. Values of $1 + C$ were used in Figure 17 to make the range comparable to that of B .

APPENDIX

A group mean performance function is a function of the same form as the function describing individuals, with mean values of the individuals' parameters, only when the function is linear in the parameters. To illustrate, suppose the performance function in Equation 2 of the text is a polynomial

$$P_{i,n} = \theta_{i1} + \theta_{i2}n + \theta_{i3}n^2 + e_i, \tag{A1}$$

where i indexes a particular subject in a group. Averaging both sides of Equation A1 over i yields for the group mean

$$M[P_n] = M[\theta_1] + M[\theta_2]n + M[\theta_3]n^2 + M[e],$$

a polynomial of similar form with group mean values for the parameters. Averaging has produced no distortion of the trend for individuals.

In contrast, suppose that the performance function for individuals is a power function

$$P_{i,n} = \theta_{i1}(n^{\theta_{i2}}). \tag{A2}$$

Averaging over both sides of Equation A2 yields for the group mean

$$M[P_n] = M[\theta_1(n^{\theta_2})],$$

and, if there is any variation in parameter values among individuals, the right side of this last expression is not equal to $M[\theta_1](n^{M[\theta_2]})$.

(Manuscript received November 19, 2001;
accepted for publication November 27, 2001.)