Psychological Review
2000, Vol. 107, No. 2, 358–367

0033-295X/00/$5.00 DOI: 10.1037//0033-295X.107.2.358

THEORETICAL NOTES

How Persuasive Is a Good Fit? A Comment on Theory Testing

Seth Roberts
University of California, Berkeley

Harold Pashler
University of California, San Diego

Quantitative theories with free parameters often gain credence when they closely fit data. This is a mistake. A good fit reveals nothing about the flexibility of the theory (how much it cannot fit), the variability of the data (how firmly the data rule out what the theory cannot fit), or the likelihood of other outcomes (perhaps the theory could have fit any plausible result), and a reader needs all 3 pieces of information to decide how much the fit should increase belief in the theory. The use of good fits as evidence is not supported by philosophers of science nor by the history of psychology; there seem to be no examples of a theory supported mainly by good fits that has led to demonstrable progress. A better way to test a theory with free parameters is to determine how the theory constrains possible outcomes (i.e., what it predicts), assess how firmly actual outcomes agree with those constraints, and determine if plausible alternative outcomes would have been inconsistent with the theory, allowing for the variability of the data.

Many quantitative psychological theories with free parameters are supported mainly or entirely by demonstrations that they can fit data—that the parameters can be adjusted so that the output of the theory resembles actual results. The similarity is often shown by a graph with two functions: one labeled *observed* (or *data*) and the other labeled *predicted* (or *theory* or *simulated*). That the theory fits data is supposed to show that the theory should be taken seriously—should be published, for example.

This type of argument is common; judging from a search of *Psychological Abstracts* (1887–1999), the research literature probably contains thousands of examples. Early instances involved sensory processes (Hecht, 1931) and animal learning (Hull, 1943), but this reasoning is now used in many areas. Here are three examples.

1. Cohen, Dunbar, and McClelland (1990) proposed a parallel distributed processing model to explain the Stroop effect and related data. The model was meant to embody a "continuous" view of automaticity, in contrast to an "all-or-none" view (Cohen et al., 1990, p. 332). The model contained many adjustable parameters, including number of units per module, ratio of training frequencies, learning rate, maximum response time, initial input weights, indirect pathway strengths, cascade rate, noise, magnitude of attentional influence (two parameters), and response-mechanism parameters (three). The model was fit to six data sets. Some parameters (e.g., number of units per module) were separately adjusted for each data set; other parameters were adjusted on the basis of one data set and were held constant for the rest. The function relating cycle time (model) to average reaction time (observed) was always linear, but its slope and intercept varied from one data set to the next. That the model could fit several data sets led Cohen et al. to conclude that, compared with the all-or-none view, "a more useful approach is to consider automaticity in terms of a continuum" (p. 357), although they did not try to fit a model based on the all-or-none view.

2. Zhuikov, Couvillon, and Bitterman (1994) presented a theory to explain avoidance conditioning in goldfish. It is a quantitative version of Mowrer's (1947) two-process theory, in which some responses are generated by fear, some by reinforcement. When some simplifying assumptions are made, the theory has three equations and six adjustable parameters. Zhuikov et al. fitted the theory to data from four experiments and concluded that "the good fit suggests that the theory is worth developing further" (p. 32).

3. Rodgers and Rowe (1993) proposed a theory that explains how teenagers come to engage in various sexual behaviors for the first time. It emphasizes contact with other teenagers—a "contagion" (Rodgers & Rowe, 1993, p. 479) explanation. The theory has eight equations with 12 free parameters. Rodgers and Rowe fitted the theory to survey data about the prevalence of kissing, petting, and intercourse in boys and girls of different ages and races and concluded that the theory "appears to have successfully captured many of the patterns in two empirical data sets" (p. 505). This success was the main support for the theory.

Seth Roberts, Department of Psychology, University of California, Berkeley; Harold Pashler, Department of Psychology, University of California, San Diego.

We thank Jonathan Baron, William Batchelder, Nicholas Christenfeld, Brett Clementz, Max Coltheart, Douglas Hintzman, James Johnston, David Krantz, James McClelland, Craig MacKenzie, Dominic Massaro, Douglas Rohrer, David Rubin, Eric Ruthruff, Saul Sternberg, James Townsend, Ben Williams, and John Wixted for their helpful comments. We also thank Saul Sternberg for providing us with his unpublished data.

Correspondence concerning this article should be addressed to Seth Roberts, Department of Psychology, University of California, Berkeley, California 94720-1650. Electronic mail may be sent to roberts@socrates.berkeley.edu.

Why the Use of Good Fits as Evidence Is Wrong

This type of argument has three serious problems. First, *what the theory predicts*—how much it constrains the fitted data—*is unclear.* Theorists who use good fits as evidence seem to reason as follows: If our theory is correct, it will be able to fit the data; our theory fits the data; therefore it is more likely that our theory is correct. However, if a theory does not constrain possible outcomes, the fit is meaningless.

A prediction is a statement of what a theory does and does not allow. When a theory has adjustable parameters, a particular fit is only one example of what it allows. To know what a theory predicts for a particular measurement, one needs to know all of what it allows (what else it can fit) and all of what it does not allow (what it cannot fit). For example, suppose two measures are positively correlated, and it is shown that a certain theory can produce such a relation—that is, can fit the data. This does not show that the theory *predicts* the correlation. A theory predicts such a relation only if it cannot fit other possible relations between the two measures (zero correlation or negative correlation), and this is not shown by fitting a positive correlation.

When a theory *does* constrain possible outcomes, it is necessary to know by how much. The more constraint—the narrower the prediction—the more impressive a confirmation of the constraint (e.g., Meehl, 1997). Without knowing how much a theory constrains possible outcomes, you cannot know how impressed to be when observation and theory are consistent.

Second, *the variability of the data* (e.g., between-subject variation) *is unclear.* How firmly do the data agree with the predictions of the theory? Are they compatible with the outcomes that the theory rules out? The more conclusively the data rule out what the theory rules out, the more impressive the confirmation. For example, suppose a theory predicts that a certain measure should be greater than zero. If the measure *is* greater than zero, the shorter the confidence interval, the more impressive the confirmation. That a theory fits data does not show how firmly the data rule out outcomes inconsistent with the theory; without this information, you cannot know how impressed to be that theory and observation are consistent.

Adding error bars may not solve this problem; it is variability on the constrained dimension or dimensions that matters. For example, suppose a theory predicts that several points will lie on a straight line. To judge the accuracy of this prediction, the reader needs to know the variability of a measure of curvature (or some other measure of nonlinearity). Adding vertical error bars to each point is a poor substitute (unless the answer, linear or nonlinear, is very clear); the vertical position of the points is not what the theory predicts.

To further illustrate these points, Figure 1 shows four ways a "two-dimensional" prediction—a constraint involving two measures at once—can be compatible with data. Measures A and B in Figure 1 are both derived from measurements of behavior. Either might be quite simple (e.g., trials to criterion) or relatively complex (the quadratic component of a fitted function); it does not matter. The axis of each measure covers the entire range of plausible values of the measure before the experiment is done (e.g., from 0 to 1, if the measure is a probability). The dotted area shows the predictions of the theory, the range of outcomes that are consistent with the theory. In the two upper panels of Figure 1, the theory tightly constrains possible outcomes; in the two lower panels, it does not. In each case, there is one data point. In the two left-hand panels, the observations tightly constrain the population value; in the two right-hand panels, they do not. In every case, the data are consistent with the theory (the data point is in the dotted area), which means that, in every case, the theory can closely fit the data. But only the situation in the upper left panel provides substantial evidence for the theory.

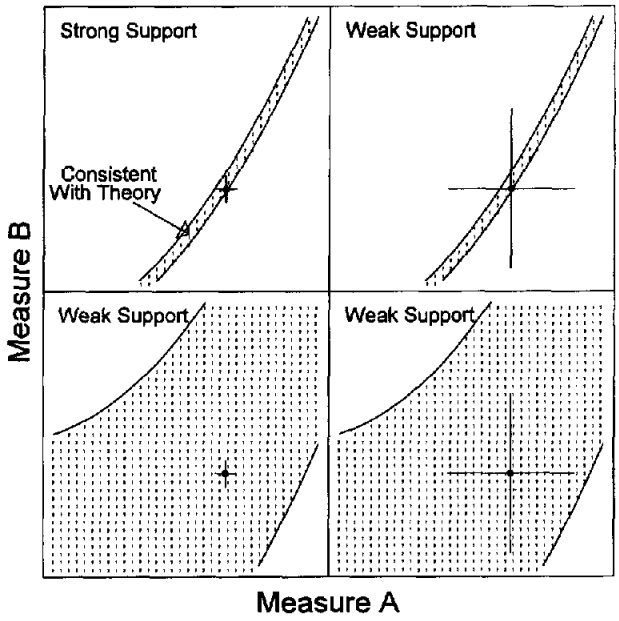

Figure 1. Four possible relationships between theory and data. Measures A and B are measures of behavior. For both measures, the axes cover the whole range of plausible values. The dotted areas indicate the range of outcomes that would be consistent with the theory. The error bars indicate standard errors. In every case, the theory can closely fit the data, but only when both theory and data provide substantial constraints does this provide significant evidence for the theory.

Third, *the a priori likelihood that the theory will fit*—the likelihood that it will fit whether or not it is true—*is ignored.* Perhaps the theory could fit any plausible result. It is well-known that a theory gains more support from the correct prediction of an unlikely event than from the correct prediction of something that was expected anyway. Lakatos (1978) made this point vividly: "It is no success for Newtonian theory that stones, when dropped, fall towards the earth, no matter how often this is repeated. . . . What really count are [the confirmation of] dramatic, unexpected, stunning predictions" (p. 6), such as the return of Halley's comet. "All the research programmes [i.e., theories] I admire have one characteristic in common. They all predict novel facts, facts which had been either undreamt of, or have indeed been contradicted [i.e., predicted to not occur] by previous or rival programmes" (Lakatos, 1978, p. 5).

Bayes's theorem, interpreted as a statement about degrees of belief, is a quantitative version of this idea. Bayes's theorem is

$$P(H|E) = \frac{P(E|H)}{P(E)} P(H),$$

where H (hypothesis) is a theory and E (event) is a particular outcome (Howson & Urbach, 1993, p. 28). P(H) is the plausibility of H before data collection, P(E) is the perceived likelihood of P before data collection, P(E|H) is the likelihood of E given that H is true, and P(H|E) is the plausibility of H after data collection—after E has been observed. When E is a prediction of H, P(E|H) = 1. Thus, according to this theorem, when P(E) is large—close to 1—observation of E will have little effect on belief in H. "Strong inference" experiments (Platt, 1964), in which different theories make contradictory predictions, are a practical application of this idea. They embody the notion that the best evidence for a theory is evidence that would be otherwise unlikely. For more discussion of the importance of the a priori likelihood of a prediction, see Howson and Urbach (1993, especially pp. 123–126).

This principle—predictions should be surprising—is relevant to psychology because psychological data are often not surprising. Therefore prediction of such data cannot provide much support for any theory. Quantitative theories are usually fit to functions: a measure of behavior (y) recorded at several values of a procedural variable (x), for example, probability of correct recall as a function of retention interval. It is never plausible that the points on the function are independent of each other, in the sense that knowing the y values of some of the points does not help you predict the y values of the rest of the points. And the lack of independence is not trivial; inevitably the plausible outcomes are a tiny fraction of the possible outcomes.

The need to make predictions that are at least a little implausible seems to have been overlooked by quantitative theorists. When a theory with three free parameters is used to fit a function with 20 data points, 20 (x, y) pairs, it is obvious that the theory must somehow constrain the function; it could not fit all possible functions with 20 points (keeping the x values fixed but allowing the y values to vary). Plainly, some results would contradict the theory. This seems to have been the sort of reasoning, either implicit or explicit, that has convinced theorists and reviewers that the data provide a test of the theory. But whether any *plausible* results would contradict the theory is not so clear.

An especially simple example of the problem involves asymptotic behavior. Suppose a learning experiment measured percent correct as a function of trial number. Performance improved for several trials but eventually—say, after 15 trials—leveled off at a value less than 100% correct—say, 93%. To fit this data, a theory will presumably need a parameter that somehow corresponds to the asymptotic level of performance (93% correct) and a parameter that corresponds to when this level is reached (after 15 trials). It needs these two adjustable parameters because both aspects of the data, 15 trials and 93% correct, surely depend on procedural details. Yet once these two parameters are properly set the theory will accurately predict performance at an unlimited number of trial numbers: It will predict 93% correct on Trial 16, on Trial 17, and so forth. If the experiment measured asymptotic performance for 50 trials (Trials 16–65), a theory—any theory—could quite accurately predict 50 data points with only two free parameters. Yet this success would add nothing to the theory's credibility.

A defender of the use of good fits as evidence might reply that fits are often judged by the percentage of variance explained, a measure that fitting the same data value (e.g., 93%) many times does not increase very much. However, the problem does not go away when the fitted data vary. The functions used to assess psychological theories are almost always "smooth," in the sense that if you know, for example, the extreme y values and alternate intermediate values (e.g., if $x = 1, 2, \ldots, 9$, the values of y for $x = 1, 3, 5, 7$, and 9), you can quite closely estimate the remaining values of y by linear interpolation. This means that any theory that does a good job of fitting about half of the data will do a good job of fitting the other half, regardless of the theory's correctness. Suppose, for example, the function consists of nine points at $x = 1, 2, \ldots, 9$. A theory with five orthogonal parameters is fit to the data for $x = 1, 3, 5, 7$, and 9, which it will fit perfectly. (The n parameters of a formula or theory are *orthogonal* if the function can fit exactly n data points. For example, $a + bx$ has two orthogonal parameters, but $ax + bx$ does not.) Then the fitted parameters are used to predict the values of y for $x = 2, 4, 6$, and 8. The theory—any plausible theory—will do a very good job. Although it is standard practice to say there were four degrees of freedom with which to test the theory, this is misleading; the goodness of fit was inevitable and therefore provides no support for the theory. The smoothness of almost any psychological function seems inevitable (i.e., is very plausible before measurement) because of both previous data (the number of smooth functions in that area of research is large and the number of jagged functions is zero or near zero) and extant theories (which predict smooth functions). If jagged functions began to be observed, or if plausible theories predicted jagged functions, then—and only then—would the prediction that a function will be smooth be interesting.

But the heart of the problem is not using constant functions or smooth functions to test theories; it is using functions that have simple shapes. Most functions measured by psychologists, and most functions to which quantitative theories are fitted, are concave up, concave down, or indeterminate between the two (i.e., close to linear). For example, learning curves (performance as a function of number of training trials) and retention functions (memory as a function of time since learning) usually fit this description. With typical amounts of data, we suspect that several equations with three orthogonal parameters, such as a quadratic equation, will fit reasonably well. The residuals may appear systematic, but the remaining structure (the structure in the residuals after the three-parameter fit is removed) will probably be impossible to detect reliably. The number of psychological research reports that have found a reliable cubic component, or reliable structure in the residuals after a three-parameter fit is removed, is very low (leaving aside psychophysical experiments and steady-state animal experiments). For indications of the typical precision of retention functions, see Rubin and Wenzel (1996) and Rubin, Hinton, and Wenzel (1999).

The practical effect of these considerations is that such functions can usually provide only a little guidance in choosing a theory, regardless of how many points they contain. The "first-degree" structure (overall level) is uninteresting; the sign of the "second-degree" structure (slope) is usually obvious (e.g., memory decays with time, performance improves with practice), and its size is uninteresting (because it presumably depends on procedural details not covered by theory); and the "fourth-degree" and higher structures cannot be made out. That leaves the "third-degree" structure (curvature) as a source of guidance. If the data were remarkably close to linear on some scale (the original y or x scales or some transformation, such as logarithmic, of either or both), that would be quite useful because most two-parameter theories would fail to

predict it (they would produce only curved functions on that scale), but that is rare. If the data were convincingly concave up (say), and this is not due to floor or ceiling effects, the best one can do is determine what sort of theories do *not* predict this, that is, what this finding rules out; perhaps it will cut the number of plausible candidate theories in half. That is progress, of course, but it cannot strongly favor any one theory. (The difficulty of extracting much information from the usual functions suggests that theorists should also look for predictions that relate two measures of behavior, as in Figure 2, presented later in this article.)

It matters that the plausible outcomes are a small fraction of the possible outcomes, because the plausible theories are crowded into the same small space, in the sense that they can predict the plausible outcomes and no others (e.g., they can predict only smooth functions). In the early days of chemistry, it was repeatedly determined that when hydrogen gas and oxygen gas combined to form water, the volume of oxygen used up was very close to half the volume of the hydrogen used up (Ihde, 1964). After several repetitions of this result, it became the only plausible, in the sense of *unsurprising,* outcome of those measurements. However, the predictions of plausible theories of the composition of water (HO? HO_2? H_2O?) remained scattered, that is, predicted a wide range of combining ratios. This is why the actual ratio could be used to choose between them. In contrast, the psychological results we have been discussing—behavior at asymptote, smooth functions, and functions with simple shapes—are both (a) likely on the basis of experience and (b) easily explained. When performance reaches asymptote and stays there (i.e., no sudden drops), we are not only not surprised but also not puzzled. It is easy to think of a theory of learning that predicts that after performance reaches asymptote it will stay there; indeed, it is hard to think of a theory that predicts anything else. When a function turns out to be smooth, it is not only unsurprising but unmysterious; it is hard to think of a theory that would not produce a smooth function. Likewise for functions with simple shapes: At the level of precision to which they are measured in most experiments, these results not only are unsurprising but could be produced by many different plausible theories.

Clearly, then, showing that a theory fits data is not enough. By itself, it is nearly meaningless. Because of the flexibility of many theories, the variability of measurements, and the simplicity of most psychological data functions, it is often quite possible that the theory could fit any plausible outcome to within the precision of the data. The reader has no way of knowing which panel of Figure 1 the evidence resembles.

Similar Criticisms

Criticisms of the use of good fits as evidence have been made by others, usually in the context of specific models (Coltheart & Coltheart, 1972; Hintzman, 1991; Johnston, van Santen, & Hale, 1985; Massaro, 1988; Roberts & Sternberg, 1993; Roediger & Neely, 1982; Wexler, 1978). When discussing specific models, these critics have often shown, or pointed out, not only that this sort of evidence may be misleading—as we argue—but that it has been misleading. These demonstrations fall into three categories.

1. *A theory "fits too much"*—it can generate such a wide range of outcomes that the fact that it can generate the actual results means little. For example, Massaro (1988) showed that "a single connectionist model can simulate results that imply [i.e., were generated by] mutually exclusive psychological processes" (p. 219). Wexler (1978), reviewing J. R. Anderson's (1976) ACT theory, noted that "ACT can model not only the Sternberg result, but also its opposite, or anything else of the sort" (p. 338). This flexibility makes the theory "so weak that there is no way to find evidence either for or against it" (Wexler, 1978, p. 346).

2. *The same data can be closely fit by a similarly flexible theory making quite different assumptions.* This means, of course, that the fits do not meaningfully support the assumptions of the theory. For example, Salasoo, Shiffrin, and Feustel (1985) found that a model with 14 free parameters could fit a variety of word-recognition data. Johnston et al. (1985), using Salasoo et al.'s data, showed that "a large family of rather different models" (p. 507) with roughly the same number of free parameters could also fit the data. Johnston et al. concluded, "Because our models fit the data [equally well] assuming only one higher level memory representation, there is no support for the assumption [of Salasoo et al.'s model] that two kinds of memories—episodic and permanent—underlie the effects of repetition on identification" (p. 507).

3. *Although a theory closely fits the data, at least one of its assumptions is wrong.* For example, as pointed out by Coltheart and Coltheart (1972), the concept-learning model of Bower and Trabasso (1964) "achieved extraordinary correspondences between predicted and obtained results" (p. 294), yet one of the assumptions of the model (independence of path) turned out to be wrong (Trabasso & Bower, 1966). In addition, Coltheart and Coltheart pointed out that four assumptions of Rumelhart's (1970) model of tachistoscopic recognition were incompatible with experimental evidence, yet the model quite closely fitted the data. Like us, Coltheart and Coltheart concluded that "it is poor strategy to evaluate a mathematical theory only by assessing how well" (p. 294) it can fit data. According to Hintzman (1991), Bower's (1961) model of paired-associates learning fitted numerous data sets with "incredible precision" (p. 50), although its central assumption was evidently quite wrong.

Although each critique (with the exception of Coltheart & Coltheart, 1972) focused on a particular theory, the diversity of the theories raises the possibility that the fundamental problem is not with any one theory or class of theories (e.g., connectionist theories are too flexible) but something broader. We suggest that the fundamental problem, as Coltheart and Coltheart argued, is a method of theory evaluation (fitting theories to data) so inadequate that serious flaws go undetected.

In the fields of statistics and computer science, a problem related to what we criticize here, called *overfitting,* has been familiar for many years (e.g., Anscombe, 1967; Leahy, 1994; Schaffer, 1993). The possibility of overfitting arises when a model that does a good job of fitting data performs poorly in other ways. For instance, a neural network program trained to classify fruit using one sample eventually achieved 90% accuracy but did much worse with a second sample from the same population (Donlin & Child, 1992). Overfitting occurs when the model is too flexible. Such experiences have taught statisticians and computer scientists that models should not be judged only by how well they fit a data set; there also must be assessment of, and penalty for, flexibility (e.g., C. M. Hurvich, 1997).

Although the arguments against the use of good fits as evidence strike us as overwhelming, we nevertheless try to present the other side—arguments in favor of the practice. In what follows, we

consider several ways in which the use of good fits as evidence might conceivably be justified—by the philosophy of science, the history of psychology, and arguments that the practice is acceptable when certain conditions are met.

Does the Philosophy of Science Support the Use of Good Fits as Evidence?

Can the use of good fits to support theories be justified by some well-accepted doctrine in the philosophy of science? Philosophers of science do not seem to have considered this particular practice, but of course much has been written about the general question of how to test theories, with considerable consensus (Howson & Urbach, 1989; Kitcher, 1993). Suppose we have a theory that we want to test. According to this consensus, there are essentially two ways to do this.

First, we can *test a prediction of the theory,* that is, make an observation that might yield results that would contradict the theory. Karl Popper, probably the most influential philosopher of science (Bondi, 1992), advocated "falsifiability" as the essential feature of scientific inquiry. According to Popper (1959), a theory must specify some possible observations that could falsify it, and a theory is supported by observations only if the observations might have had outcomes inconsistent with the theory.

Second, if there are competing (incompatible) explanations of the facts that our theory explains, we can *test a prediction of a competing theory.* In many cases, alternative theories are incompatible; that is, if one theory (T_0) is correct, other explanations (T_1, T_2, etc.) of the same facts must be wrong. In these cases, elimination of alternatives supports T_0. This approach was first sketched by Bacon (1620/1960; Urbach, 1987).

If alternative theories exist and make differing predictions (e.g., one theory says a certain measurement should be zero, whereas another theory says it should be positive), we can combine the two approaches and *test a prediction of the theory and a prediction of a competing theory at the same time.* When the two predictions are incompatible (nonoverlapping), this is what Platt (1964) called *strong inference.* (*Efficient* inference may have been a better name. The results will not be decisive—"strong"—unless several other conditions are met.)

When it is claimed that a good fit supports a theory, what sort of test is this? Nothing is said about competing theories, eliminating the second method. Perhaps theorists who support theories with good fits to data believe that they have tested a prediction of the theory (the prediction that the theory will fit the data), a Popperian test. But they have not shown that, given the precision of the data, there were any plausible outcomes that the theory could not have fit.

Thus we did not find any support within the philosophy of science for the use of good fits to support theories.

Does the History of Psychology Support the Use of Good Fits as Evidence?

The use of close fits as evidence might be justified by showing that it has "worked"—led to demonstrable progress. We searched the history of psychology for theories originally supported mainly or entirely by good fits to data that eventually found support from other sources (e.g., tests of specific assumptions, confirmation of new predictions). We were unable to find even one example. Although several reviewers of this article disagreed with our conclusions, they did not provide examples of such a theory.

An early example of the use of close fits by themselves to support a theory is Hecht's (1931) theory of color vision—a theory that is almost completely forgotten nowadays. In contrast, Hering's (1878/1964) theory of color vision, based on quite different data, is still important (L. M. Hurvich, 1981). Another early example of the practice is *Principles of Behavior* (Hull, 1943), which may have been cited more often than any other work in the experimental psychology literature of the 1940s and 1950s. In spite of numerous excellent fits, it seems fair to say that none of Hull's theoretical ideas supported by fitted curves are still influential. Mackintosh (1983), for instance, referred to the "legacy" (p. 2) of Thorndike, Pavlov, Konorski, and Tolman, but not Hull.

Later quantitative learning theories were much simpler than Hull's (1943) but still relied on good fits for support. In what Jenkins (1979) called a "ground-breaking paper" (p. 206), Estes (1950) used the following equation, derived from a theory of learning, for the mean latency to fit some runway learning data, with L indicating the latency to leave the start box and T the trial number:

$$\bar{L} = \frac{2.5}{1 - .9648e^{-.127T}}.$$

The parameters 2.5, .9648, and −.12 were of course estimated from the data. According to Estes, the fit was "satisfactory" (p. 101). Satisfactory or not, a reader could not know what to make of this evidence. The variability of the data was not shown, so it was unclear if the deviations were reliable. Nor was it clear whether any plausible results could have contradicted the theory. Although many theorists seemed to have been impressed at the time—as Jenkins said, Estes's work led to many similar theories—later theorists were less impressed. A look at any recent text on animal learning suggests that the mathematical learning theorists of the 1950s and 1960s, in spite of many successful fits, discovered nothing that formed the basis for current theories of learning.

The use of good fits as evidence probably received a boost from the advent of cheap and powerful computers, which made it much easier to search a large parameter space for the best fit. Connectionist theorizing, in particular, took advantage of the new flexibility in model building that seemed to be available. An influential early article in this area (J. A. Anderson, 1973) proposed an explanation for some reaction time results with short memorized lists. Empirical support for the theory consisted almost entirely of demonstrations that it could fit a variety of data. The fits involved five to eight free parameters, which changed from one data set to the next. It was unclear what the theory predicted, that is, what it could not fit; because the constraints were unclear, variability on the constrained dimensions was also unclear. Because the number of data points was much larger than the number of free parameters, the theory surely ruled out many possible outcomes, but whether it ruled out any plausible outcomes was not clear.

An example of later research along these lines is Seidenberg and McClelland's (1989) theory of visual word recognition and pronunciation. Their goal was a connectionist model "that exhibited many of the basic phenomena of word recognition and naming" (Seidenberg & McClelland, 1989, p. 529). The evidence for the

model consisted of numerous graphs that showed a close fit between two measures: reaction time (observed in experiments) and squared error (produced by the model). What the model could not fit was unclear.

In Hinton and Anderson (1981) and Rumelhart, McClelland, and the PDP Research Group (1986), the first influential books on connectionism, the issue of how to test such flexible theories received almost no attention. In spite of the popularity of connectionist models, and numerous good fits, we have yet to encounter even one such model whose predictions have been determined, much less verified or shown to rule out plausible results. Massaro (1988) made similar points. Without accurate predictions in cases in which the prediction could have plausibly been wrong, the claim that connectionist theories have helped us understand the brain seems to rest entirely on belief in the assumptions of these theories.

So we did not find any support in the history of psychology for the use of good fits to support theories.

Defenses of the Use of Good Fits as Evidence

Many psychologists, we suspect, realize that not all good fits provide substantial support for a theory. Yet they believe that *their* example is sound because it satisfies certain conditions. Although the use of good fits as evidence may in general be flawed, they believe that in certain restricted situations it is helpful. Here we consider the arguments along these lines that we have encountered most frequently.

Defense 1: A good fit is impressive when there are more observations in the data set than free parameters in the model. "A standard rule of thumb states that a model has too many [free] parameters to be testable if and only if it has at least as many parameters as empirically observable quantities" (Bamber & Van Santen, 1985, p. 443). For example, if a model has five free parameters and there are 20 data points, there supposedly are 15 degrees of freedom for assessing the fit.

It is a generous rule of thumb. In fact, the number of free parameters in a theory provides an upper bound on its flexibility. If a theory has five orthogonal free parameters, then it will be able to fit exactly any five data points; if the parameters are not orthogonal, however, the number of data points the theory can fit exactly is less (as in the example given earlier, $ax + bx$, which has two parameters, a and b, but cannot fit any two data points). The more serious distortion, however, is the idea that the number of data points indicates the range of possible outcomes—that if there are 10 data points, the possible outcomes could have plausibly been anywhere in a 10-dimensional space. As we argued above, this is usually a great overstatement. A more realistic view is that most functions provide only one useful piece of information for testing theories: whether the function is concave up, nearly linear, or concave down (when the data are scaled so that all three possibilities are plausible).

Defense 2: My model fits better than another model. Theorists often compare the fits produced by different models and assume that the best fitting one deserves belief because it has won a kind of competition (e.g., Ashby & Lee, 1991; Atkinson & Crothers, 1964; Bush & Mosteller, 1959; Nosofsky, Kruschke, & McKinley, 1992). There are several problems with this approach. First, the best fitting model may merely be the most flexible model rather than the best model (Collyer, 1985)—a lesson that statisticians and computer scientists learned long ago, as discussed above. To equate the flexibility of the theories being compared, psychologists sometimes adjust goodness-of-fit statistics according to a general formula (Akaike, 1974; Takane & Shibayama, 1992). Unfortunately, this method is inadequate because the flexibility added by a free parameter depends on the details of the theory (cf. $ax + bx$ with $ax + b$; both have two parameters, but the latter is more flexible). The only accurate way to "allow" for the flexibility of a theory, as far as we know, is to determine what the theory predicts. Second, assuming the best fitting model is best takes no account of the variability of the data. Suppose, for example, that Theory X predicts that a certain measurement should be 5 whereas Theory Y predicts that it should be 7. If the actual result is 5.5 ± 10, Theory X will fit better, yet there is no good reason to prefer it.

Fitting several plausible models to learn if any can be *ruled out* makes sense, especially when combined with an effort to find features of the data that are hard to fit. But this is not what is usually done. For example, Rodgers and Rowe (1993), in their study of teenagers' sexual behavior, fitted two different models making somewhat different assumptions. Although "both models were consistent with the data according to chi-square tests" (p. 495), Rodgers and Rowe favored one of them.

Comparing the fit of several theories should not be confused with comparing their predictions, which is always worthwhile. In these fit-comparison situations, the predictions—that is, the constraints—of the various theories are not even determined, much less compared, at least in the examples we have seen.

Defense 3: The research and editorial processes protect readers from too flexible models. During the theory-building process, the argument goes, many models are rejected because they cannot fit the data. When the theorist finally finds a model that *can* fit the data, he or she hurries to publish it and does not describe in the publication all the failures. A problem with this argument is that a reader has no way of knowing if it is true; nor can the reader be sure that the published theory is no more flexible than the rejected theories. A similar argument is that reviewers can supposedly tell when a model is too flexible. Again, a reader has no way of knowing if this is true. The plausibility of contradictory outcomes, outcomes that the theory cannot fit, is crucial information that should be made explicit.

Better Ways to Judge Theories With Free Parameters

The problems described earlier have straightforward solutions.

Problem 1: What the theory predicts is unclear. Solution: Determine the predictions. To determine the predictions of a theory with free parameters requires varying each free parameter over its entire range, in all possible combinations (i.e., surveying the entire parameter space). For each combination of parameters (each point in the parameter space), the theory generates simulated behavior. The prediction of the theory, for any measure (any function of the observations, real or simulated), is the range of outcomes that the theory can produce (Sternberg, 1963, pp. 89–90). For example, suppose a theory has two free parameters: a, which can range from 0 to 10, and b, which can range from 0 to 1. To determine what the theory predicts for, say, trials to criterion, one would vary both a and b over their entire ranges, in all possible combinations (i.e., over the whole two-dimensional parameter space), and determine the predicted trials to criterion for each combination of

parameter values (i.e., for each point in the parameter space). The prediction of the theory for this dimension of data would be the entire range of trials to criterion that the theory could produce. Using intuition, experience, and trial and error, the theorist must search among the many predictions of a theory to find those narrow enough to plausibly be falsified.

Problem 2: The variability of the data is unclear. Solution: Show the variability of the data. As discussed above, it is variability on the constrained dimensions that is important. This means that Problem 1 (unclear predictions) must be solved first.

Solutions to Problems 1 and 2 are illustrated by Roberts and Sternberg (1993), who tested Ashby's (1982) version of McClelland's (1979) cascade model. The tested version of Ashby's model had two free parameters: the time constants of two processes. Roberts and Sternberg varied those parameters over all plausible values they could have in a 2 × 2 experiment. Examination of simulated results covering the entire parameter space showed that a certain measure derived from reaction times (a *main effect difference statistic*) was constrained by the model and that this constraint varied with a second measure (a *variance-change statistic*). Both statistics vaguely resemble measures of interaction. Figure 2, from Roberts and Sternberg, shows this prediction and some data. Each small point represents the results of a simulated 2 × 2 experiment; the area filled by these points is the prediction of the theory. The large points, with standard error bars based on between-subjects variation, represent data. Some of the points fall within the predicted area but none firmly, and several points fall firmly outside the predicted area, which is inconsistent with the model. Thus, the model fails the test.

Problem 3: Perhaps the theory could fit any plausible result. Solution: Show that there are plausible results the theory cannot fit. It is not enough to show that there are some results the theory cannot fit. To meaningfully constrain the data, there must be some *plausible* results the theory cannot fit.

Suppose we were to test a theory and discover that it accurately predicts the results, that is, theory and data are consistent. Which quadrant of Figure 1 does the evidence resemble? To find out, we would need to determine the range of plausible alternative results—predictions different from the prediction of the theory being tested. How we decide what is plausible is a big subject (e.g., Hogarth, 1980), but everyone agrees that both theory (beliefs about how the world works) and data (actual observations) are important, that we use both to judge the likelihood of future events. For example, Lakatos (1978), in the statement quoted earlier, mentioned both. It is important, he said, that predictions be surprising—that they differ from "stones falling to earth when dropped" (expectations based on experience) or from expectations based on "rival programmes" (predictions of other theories).

Determining what other theories predict needs no explanation. However, the idea of determining what experience (unexplained by any theory) predicts may be unfamiliar. Earlier measurements similar to the current measurement may have generated a range of outcomes, which would suggest that the current measurement could have a similar range. Or earlier measurements may have

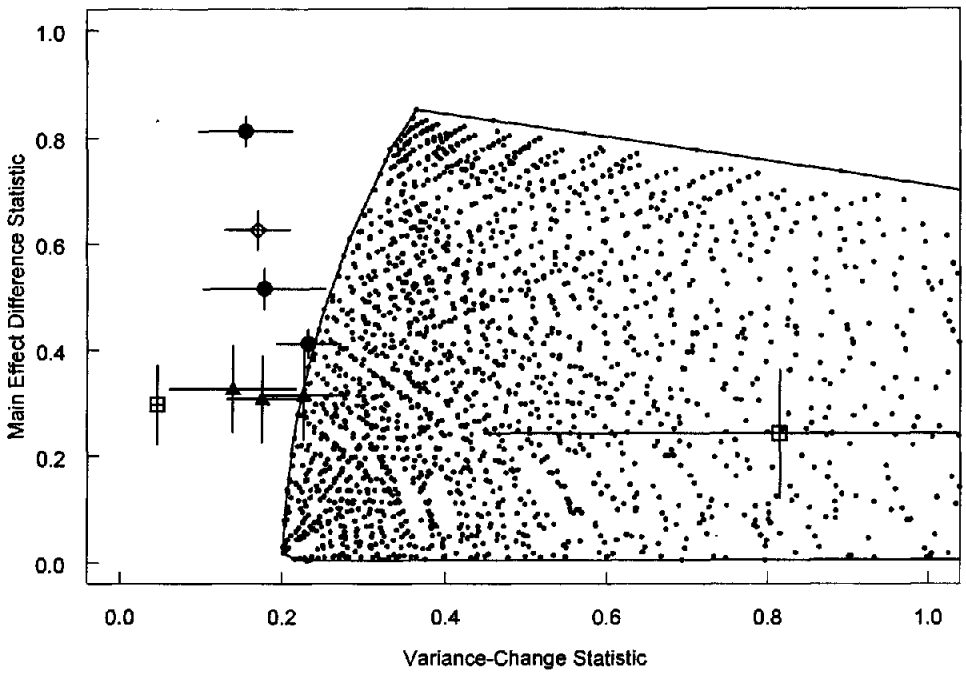

Figure 2. A prediction of a version of Ashby's (1982) cascade model and some data. Each of the many small points is derived from the results of a simulated 2 × 2 experiment. The large points, with standard error bars, are from actual experiments. From "The Meaning of Additive Reaction-Time Effects: Tests of Three Alternatives," by S. Roberts and S. Sternberg, 1993, in D. E. Meyer and S. Kornblum (Eds.), *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience—A Silver Jubilee* (p. 641), Cambridge, MA: MIT Press. Copyright 1993 by MIT Press. Adapted with permission.

suggested empirical generalizations that predict a specific value or range of values in the current case.

The range of plausible outcomes is the union of the predictions based on other plausible theories and expectations based on other data. For example, if other theories suggest the measurement might be 10 to 30, and other data suggest it might be 20 to 50, the plausible range is 10 to 50. For the observed consistency of theory and data to be meaningful, it is necessary only that *some* of this range falls outside of what the tested theory predicts. Of course, the more of this range that the tested theory cannot explain, the more impressive the observed consistency. Because pointing out plausible alternatives is rare, many theorists may not have realized that doing so would strengthen the case for the theory they favor.

To compare plausible alternative outcomes with what the tested theory could explain, it is necessary to combine (a) the flexibility of the tested theory and (b) the variability of the actual results. As Figure 1 illustrates, the evidence will not be convincing if either is large compared with the range of plausible outcomes.

In practice, this comparison requires four steps. First, *determine what the theory of interest predicts.* For example, suppose it predicts that the measurement will be between 40 and 50. Second, *determine the 95% confidence interval based on the data.* Suppose the confidence interval is the average ± 10. Third, *widen the prediction interval appropriately.* In the example, the widened interval is 30 (40 − 10) to 60 (50 + 10). The new interval (30 to 60) is the range of results (i.e., averages) consistent with the theory, given the variability of the data. Unlike familiar intervals, the actual result will probably not be in the middle of the interval. Fourth, *compare actual and plausible results with the widened interval.* The results should increase belief in the theory only if the actual result is within the widened interval and at least one plausible alternative result is outside the widened interval.

Figure 3 shows two examples. The solid line shows what the theory predicts; the dotted lines extend the prediction to allow for the variability of the data. In both cases, the tested theory could closely fit the result, but only the left-hand pattern of results should increase belief in the theory.

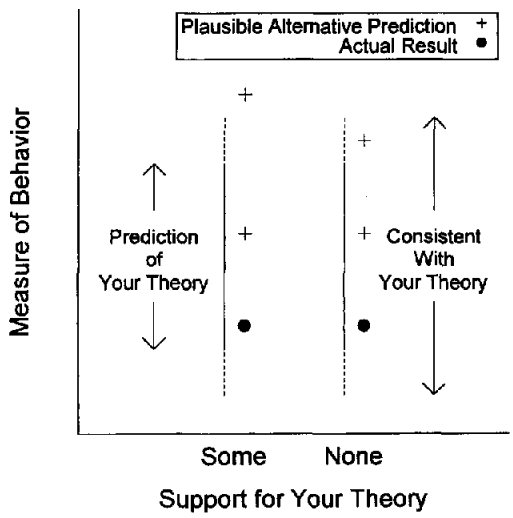

Figure 3. How the plausibility of other results affects the interpretation of the observed results. The solid lines indicate the prediction of the tested theory. The dotted lines, which are based on the variability of the data, indicate 95% confidence intervals.

Sternberg's (1966) memory-scanning data allow a simple real-life illustration. In the varied-set procedure, the participant saw a list of one to six digits. After a brief delay, the participant saw a test digit and indicated as quickly as possible whether it was on the list. The measure of interest was the reaction time to respond "yes" or "no." Mean reaction time increased with list length. An interesting theoretical question is whether the results support a theory of serial memory scanning, a simple version of which implies that the increase should be linear with list length.

All possible outcomes were not equally plausible, of course. On the basis of previous results, it was quite likely, before the experiment was done, that reaction time would change monotonically with list length—for example, that the reaction time for a list length of two would be between the reaction time for a list length of one and the reaction time for a list length of three (within experimental error). This restriction should be taken into account when one is deciding how impressed to be with observed linearity—or, more precisely, a failure to reject the hypothesis of linearity—because a large fraction of the results that would have rejected that hypothesis were implausible. To not take this into account would give the hypothesis of linearity an undeserved boost.

A realistic assessment of the evidence for linearity thus requires a plausible alternative prediction (or range of predictions). One alternative is provided by the empirical generalization that reaction time is linear with the logarithm of the number of stimulus–response combinations (Hick, 1952; Hyman, 1953). Considering each stimulus–response combination as one item (or two items) to be remembered suggests the empirical generalization that reaction time is linear with the logarithm of the number of items to be remembered. This generalization might be wrong, of course, but before Sternberg (1966) collected his data, it was plausible and therefore could be used to generate plausible outcomes. In Sternberg's experiment—assuming that each digit to be remembered is an item—it implies that reaction time would be linear with the logarithm of list length (the number of digits to be remembered). Certain theories also suggest this relation (Sternberg, 1966).

When at least one plausible alternative to linearity has been identified, it becomes possible to assess how much results consistent with linearity support a theory that predicts linearity. One way to test the prediction of linearity is to use the reaction times with lists of lengths one and six to predict by interpolation the average reaction time with lists of lengths three and four. The logarithmic prediction can be tested in a similar way. Figure 4 shows the results of this analysis. The results agree with the linear prediction but reliably differ from the logarithmic prediction. Because the results rule out a plausible alternative, the fact that they are consistent with a prediction of the serial-scanning theory provides real support for that theory.

Why Has the Use of Good Fits as Evidence Persisted?

Why has the practice of using good fits to support theories been so popular? Its flaws—it hides the flexibility of the theory and the variability of the data and ignores the plausible range of the data—are large and easy to understand. There are several possible reasons.

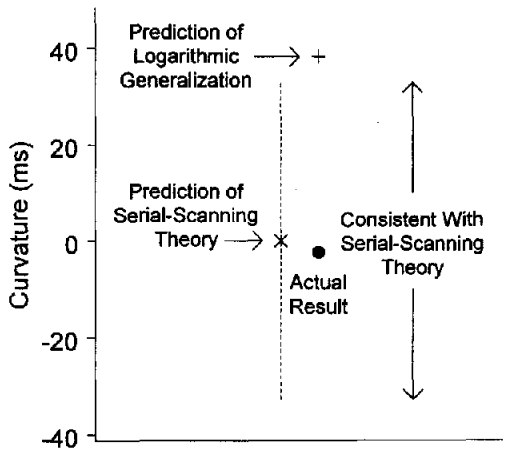

Figure 4. Assessment of a prediction of a serial-scanning theory (based on unpublished data from an experiment reported by Sternberg, 1966). Points are means; the dotted line shows a 95% confidence interval based on between-subject variance. The serial-scanning theory predicts that reaction time will be linear with list length. The alternative prediction is that reaction time will be linear with the logarithm of list length.

1. *A desire to imitate physics.* This may have been important initially. In 1929, Clark Hull "purchased and became deeply familiar with Newton's *Principia*, a work which strongly influenced his thinking from that time on" (Beach, 1959, pp. 128–129). Presenting a graph with several data points and a line through the points makes it appear that the theory being fit makes narrow quantitative predictions, like many physical theories.

2. *Confirmation bias* (J. Palmer, personal communication, November 1, 1996). Confirmation bias is a tendency to test beliefs in ways likely to confirm them. To regard a good fit as substantial evidence is of course to adopt a testing strategy that tends to confirm flexible theories. Nickerson (1998) concluded that "a great deal of empirical evidence supports the idea that the confirmation bias is extensive and strong and that it appears in many guises" (p. 177); he described several examples involving scientific practice. In many theoretical publications, the authors test only one theory—a theory that they created and that, naturally, they wish to confirm.

3. *Repetition.* Once a new result or method has appeared in print a few times, it gains a certain respect, and a certain momentum, unrelated to merit. Sheer repetition—if it is repetition of a mistake—can be strong enough to push whole scientific fields off track for many years, which is what we claim happened here. A famous example in physics involves the charge on the electron. In 1909, when Millikan measured this quantity for the first time, he used a wrong value for the viscosity of air. Subsequent measurements of the charge on the electron shifted only gradually from Millikan's value to the correct value (Feynman, 1985). Biology provides another example. From the 1930s until 1955, mammalian cytologists were "virtually certain" (Kottler, 1974, p. 465) that human cells contained 48 chromosomes, although the correct number is 46. This conclusion was based on "chromosome counts made during the 1920s and 1930s by a number of esteemed cytologists all over the world" (Kottler, 1974, p. 465). By 1954, the existence of 48 human chromosomes was "an established fact" (Kottler, 1974, p. 466), according to one cytologist. The correct number was discovered only when improved techniques made counting chromosomes much less error-prone (Kottler, 1974). Similarly, the use of good fits as evidence in experimental psychology may have remained popular at least partly because of repetition and inertia.

4. *Theory complexity.* As theories have grown in complexity, it has become no easy task to determine how they constrain possible outcomes. It is computationally much easier to fit them to data.

5. *Neglect of basic principles.* The most basic principles of theory testing—the ideas that (a) to test a theory, you must collect data that could plausibly disprove it and (b) the more plausible the possibility of disproof, the stronger the test—receive little attention in psychology. They are far from obvious; as Lakatos (1978) pointed out, Popper himself failed to appreciate the crucial role of plausibility.

A larger lesson of this article may be that these principles—and the related questions of "what would disprove my theory?" and "what theories do these data rule out?"—deserve more emphasis.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716–723.

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

Anderson, J. R. (1976). *Language, memory, and thought.* Hillsdale, NJ: Erlbaum.

Anscombe, F. J. (1967). Topics in the investigation of linear relations fitted by the method of least squares (with discussion). *Journal of the Royal Statistical Society, Series B, 90,* 1–52.

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

Ashby, F. G., & Lee, W. W. (1991). Predicting similarity and categorization from identification. *Journal of Experimental Psychology: General, 120,* 150–172.

Atkinson, R. C., & Crothers, E. J. (1964). A comparison of paired-associate learning models that have different acquisition and retention axioms. *Journal of Mathematical Psychology, 2,* 285–315.

Bacon, F. (1960). *The new organon and related writings.* New York: Liberal Arts Press. (Original work published 1620)

Bamber, D., & Van Santen, J. P. H. (1985). How many parameters can a model have and still be testable? *Journal of Mathematical Psychology, 29,* 443–473.

Beach, F. A. (1959). Clark Leonard Hull: May 24, 1884–May 10, 1952. *Biographical Memoirs, National Academy of Sciences, 33,* 125–141.

Bondi, H. (1992). The philosopher for science. *Nature, 358,* 363.

Bower, G. H. (1961). Application of a model to paired-associate learning. *Psychometrika, 26,* 255–280.

Bower, G., & Trabasso, T. (1964). Concept identification. In R. C. Atkinson (Ed.), *Studies in mathematical psychology* (pp. 32–94). Stanford, CA: Stanford University Press.

Bush, R. R., & Mosteller, F. (1959). A comparison of eight models. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory* (pp. 293–307). Stanford, CA: Stanford University Press.

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

Collyer, C. E. (1985). Comparing strong and weak models by fitting them to computer-generated data. *Perception & Psychophysics, 38,* 476–481.

Coltheart, M., & Coltheart, V. (1972). On Rumelhart's model of visual information-processing. *Canadian Journal of Psychology, 26,* 292–295.

Donlin, M., & Child, J. (1992). Is neural computing the key to artificial intelligence? *Computer Design, 31*(10), 87–100.

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review, 57,* 94–107.

Feynman, R. P. (1985). *Surely you're joking Mr. Feynman.* New York: Norton.

Hecht, S. (1931). The interrelations of various aspects of color vision. *Journal of the Optical Society of America, 21,* 615–639.

Hering, E. (1964). *Outlines of a theory of the light sense* (L. Hurvich & D. Jameson, Trans.). Cambridge, MA: Harvard University Press. (Original work published 1878)

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology, 4,* 11–26.

Hinton, G. E., & Anderson, J. A. (1981). *Parallel models of associative memory.* Hillsdale, NJ: Erlbaum.

Hintzman, D. L. (1991). Why are formal models useful in psychology? In W. E. Hockley & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock* (pp. 39–56). Hillsdale, NJ: Erlbaum.

Hogarth, R. (1980). *Judgment and choice.* New York: Wiley.

Howson, C., & Urbach, P. (1993). *Scientific reasoning: The Bayesian approach.* La Salle, IL: Open Court Press.

Hull, C. L. (1943). *Principles of behavior.* New York: Appleton-Century.

Hurvich, C. M. (1997). Mean square over degrees of freedom: New perspectives on a model selection treasure. In D. R. Brillinger, L. T. Fernholz, & S. Morgenthaler (Eds.), *The practice of data analysis: Essays in honor of John W. Tukey* (pp. 203–215). Princeton, NJ: Princeton University Press.

Hurvich, L. M. (1981). *Color vision.* Sunderland, MA: Sinauer.

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology, 45,* 188–196.

Ihde, A. J. (1964). *The development of modern chemistry.* New York: Harper & Row.

Jenkins, H. M. (1979). Animal learning and behavior theory. In E. Hearst (Ed.), *The first century of experimental psychology* (pp. 177–230). Hillsdale, NJ: Erlbaum.

Johnston, J. C., van Santen, J. P. H., & Hale, B. L. (1985). Repetition effects in word and pseudoword identification: Comment on Salasoo, Shiffrin, and Feusteel. *Journal of Experimental Psychology: General, 114,* 498–508.

Kitcher, P. (1993). *The advancement of science: Science without legend, objectivity without illusions.* New York: Oxford University Press.

Kottler, M. J. (1974). From 48 to 46: Cytological technique, preconception, and the counting of human chromosomes. *Bulletin of the History of Medicine, 48,* 465–502.

Lakatos, I. (1978). *The methodology of scientific research programmes.* Cambridge, England: Cambridge University Press.

Leahy, K. (1994). The overfitting problem in perspective. *AI Expert, 9*(10), 35–36.

Mackintosh, N. J. (1983). *Conditioning and associative learning.* Oxford, England: Oxford University Press.

Massaro, D. W. (1988). Some criticisms of connectionist models of human performance. *Journal of Memory and Language, 27,* 213–234.

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

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 393–425). Mahwah, NJ: Erlbaum.

Mowrer, O. H. (1947). On the dual nature of learning—A reinterpretation of "conditioning" and "problem-solving." *Harvard Educational Review, 17,* 102–148.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2,* 175–220.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 211–233.

Platt, J. R. (1964, October 16). Strong inference. *Science, 146,* 347–353.

Popper, K. (1959). *The logic of scientific discovery.* London: Hutchinson.

Roberts, S., & Sternberg, S. (1993). The meaning of additive reaction-time effects: Tests of three alternatives. In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience—A silver jubilee* (pp. 611–653). Cambridge, MA: MIT Press.

Rodgers, J. L., & Rowe, D. C. (1993). Social contagion and adolescent sexual behavior: A developmental EMOSA model. *Psychological Review, 100,* 479–510.

Roediger, H. L., & Neely, J. H. (1982). Retrieval blocks in episodic and semantic memory. *Canadian Journal of Psychology, 36,* 213–242.

Rubin, D. C., Hinton, S., & Wenzel, A. (1999). The precise time course of retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1161–1176.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review, 103,* 734–760.

Rumelhart, D. E. (1970). A multicomponent theory of the perception of brief visual displays. *Journal of Mathematical Psychology, 7,* 191–218.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition.* Cambridge, MA: MIT Press.

Salasoo, A., Shiffrin, R. M., & Feustel, T. C. (1985). Building permanent memory codes: Codification and repetition effects in word identification. *Journal of Experimental Psychology: General, 114,* 50–77.

Schaffer, C. (1993). Overfitting avoidance as bias. *Machine Learning, 10,* 153–178.

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

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

Sternberg, S. (1966, August 5). High-speed scanning in human memory. *Science, 153,* 652–654.

Takane, Y., & Shibayama, T. (1992). Structures in stimulus identification data. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 335–362). Hillsdale, NJ: Erlbaum.

Trabasso, T., & Bower, G. (1966). Presolution dimensional shifts in concept identification: A test of the sampling with replacement axiom in all-or-none models. *Journal of Mathematical Psychology, 3,* 163–173.

Urbach, P. (1987). *Francis Bacon's philosophy of science.* La Salle, IL: Open Court Press.

Wexler, K. (1978). A review of John R. Anderson's *Language, Memory, and Thought. Cognition, 6,* 327–351.

Zhuikov, A. Y., Couvillon, P. A., & Bitterman, M. E. (1994). Quantitative two-process analysis of avoidance conditioning in goldfish. *Journal of Experimental Psychology: Animal Behavior Processes, 20,* 32–43.

Received April 10, 1996
Revision received June 16, 1999
Accepted June 18, 1999 ■