1 Estes' Statistical Learning Theory: Past, Present, and Future

Patrick Suppes
Stanford University

THE PAST

The direct lineage of statistical learning theory began in 1950 with the publication in *Psychological Review* of Estes' article "Toward a statistical theory of learning." Before saying more about that I recall, however, that there were a number of earlier attempts to develop a quantitative theory of learning which can easily be cited, but I hasten to say that I am not attempting anything like a serious history of the period prior to 1950. I have used the following highly selective but useful procedure of listing the earlier book-length works referred to in the spate of important papers published between 1950 and 1953, which I discuss in a moment. The earlier references referred to (not counting a large number of articles) are Skinner's *Behavior of Organisms*, published in 1938, Hilgard and Marquis' *Conditioning and Learning*, published in 1940, Hull's *Principles of Behavior*, published in 1943, and Wiener's *Cybernetics*, published in 1948. These and other earlier works not mentioned played an important role in developing the theoretical concepts of learning used in statistical learning theory and related theories developed by others shortly thereafter. The basic concepts of particular importance are those of association and reinforcement expanded into the concepts of stimulus, response, and reinforcement, and the processes of stimulus sampling and stimulus conditioning. Other important psychologists who contributed to this development and who were not mentioned on the basis of the principle of selection I used are Guthrie, Thorndike, Thurstone, Tolman, and Watson.

The work of all these psychologists would seem to be almost a necessary preliminary for detailed quantitative and mathematical developments. (Wiener's

work, although mathematical in content, is not really in the tradition of psychological research.) The central contribution of statistical learning theory is the use of the psychological concepts of association and reinforcement to develop a genuinely quantitative theory of behavior. When I say *genuinely quantitative*, I have in mind something rather specific. Earlier attempts at quantitative theory, perhaps especially Hull's, did not lead to a theory that was mathematically viable. In Hull's theory it is impossible to make nontrivial derivations leading to new quantitative predictions of behavior. Hull's heart was certainly in the right place and he performed a real service in the kind of efforts he undertook, but it must be said that they were not in any deep sense successful. In contrast, beginning with Estes' 1950 paper, there was a rapid development of ideas that had the same sort of feel about it that the development of ideas and theories have in physics. This is not meant to make any lame comparison between physics and psychology, but rather to make the point that mathematically formulated theory in science, whether it be physics, psychology, or any other discipline, must be set up in such a way that nontrivial quantitative predictions can be made, which can themselves be checked by new experiments or new observations, without some unreasonable number of parameters left to be estimated for each new situation. The many experiments conducted by Estes, his colleagues, and his students over the decade and a half after 1950 attest to the ability to adapt the theory in a fruitful and interesting way to many different experimental configurations.

I turn now to that beginning of the new era in learning that I properly date with Estes' 1950 paper. The opening sentence sets the tone of this new beginning:

> Improved experimental techniques for the study of conditioning and simple discrimination learning enable the present-day investigator to obtain data which are sufficiently orderly and reproducible to support exact quantitative predictions of behavior.

By the end of 1953 a number of important theoretical articles had appeared that set the tone for another decade. I mention especially and in chronological order the two *Psychological Review* articles in 1951 of Bush and Mosteller, which presented in clear and workable mathematical fashion models of simple learning and models of stimulus generalization and discrimination; the 1952 article by George Miller on finite Markov processes; and the 1952 article by Miller and McGill on verbal learning. I end with the 1953 *Psychological Review* article by Estes and Burke on the theory of stimulus variability. These are not the only articles on learning that appeared during this period, nor even the only theoretical ones. But they are by the psychologists who created a new theoretical approach in psychology. To an unusual degree, Estes' 1950 article marks in a very definite way the beginning of this new era.

My own involvement with Bill Estes and learning theory began after 1953. We first began to work together in 1955 when we were both fellows at the Center

for the Advanced Study in the Behavioral Sciences at Stanford. It is probably fair to say that I learned more about learning theory in the fall of 1955 in my almost daily interactions with Bill at the Center than at any other time in my life. He brought me quickly from a state of ignorance to being able to talk and think about the subject with some serious understanding. I had previously tried to read the works of Hull mentioned earlier, but had found it fairly hopeless in terms of extension and further development, for the reasons already indicated. Once Bill began explaining to me the details of statistical learning theory I took to it like a duck to water, because I soon became convinced that here was an approach that was both experimentally and mathematically viable.

Over the years I have become involved in other aspects of mathematical psychology, especially the theory of measurement, the theory of decision making, and the learning of school subjects such as elementary mathematics. In the process I have had the opportunity to work with colleagues who have also taught me a lot, particularly Duncan Luce, Dave Krantz, Amos Tversky, Dick Atkinson, and Henri Rouanet. But although many years have passed since Bill and I worked together on a regular basis, I still remember vividly my own intellectual excitement as we began our collaboration in that year at the Center.

More generally in the 10 years following that initial burst of activity from 1950 to 1953, Bill had many good students and much good work was published both by him, his colleagues, and his students. In fact, in my own judgment the decade and a half from 1950 to 1965 will be remembered as one of those periods when a certain new approach to scientific theory construction in psychology had a very rapid and intense development, perhaps the most important in this century.

THE PRESENT

I think it is fair to say that in the 1970s and 1980s much of Bill's own interests shifted from learning theory to problems of memory and related areas of cognition. During this period the new wave of enthusiasm was for cognition rather than for learning. Fortunately, in the last few years we have had a return to a central concern with learning in the recent flourishing of connectionism and the widespread interest in neural networks. It is apparent enough that many of the new theoretical ideas have a lineage that goes back to the developments in learning theory between 1950 and 1965.

I am not going to try to give a survey of these new ideas (see Suppes, 1989), but rather give three applications that arise out of statistical learning theory and related work on stochastic models of learning in the 1950s. These applications are also not meant to be a survey, but reflect my own current work. My purpose in discussing them is to show how the basic ideas of statistical learning theory remain important and have the possibility of application in quite diverse areas of theoretical and practical concern.

My first example concerns the learning of elementary mathematics. The work I report here has been done with Mario Zanotti and will be written up in detail elsewhere. The second example concerns robots that learn and the work I report has been done in collaboration with Colleen Crangle. The third example concerns the learning of Boolean functions, with particular attention to the introduction of hierarchies in such learning. This work has been done with Shuzo Takahashi.

Learning Elementary Mathematics

What I describe under this heading is extensive work at Computer Curriculum Corporation over the past decade on computer-assisted instruction in basic skills, especially elementary mathematics. The fundamental problem is the continual dynamic assessment of mastery on the part of the student. The use of computers for instruction permits deep concern for operational individualization of instruction. This means that each student can be moved forward on an individual basis as he or she masters successive skills or concepts. The course entitled "Mathematics Concepts and Skills" is aimed at supplementary instruction, ranging from kindergarten to the eighth grade. The course is divided into 16 content strands. A strand itself is ordered into a sequence of equivalence classes of exercises of increasing difficulty. An equivalence class is meant to be a collection of exercises of essentially the same difficulty. For example, an equivalence class in the addition strand might contain exercises of adding two three-digit numbers with no carry from one digit to another. There are also strands concerned with measurement, geometry, word problems, and so on. I do not attempt to give here a systematic description of the content of the course, which in a broad sense overlaps extensively with any standard textbook series. The features of individualization and the organization of a given strand into an ordered sequence of equivalence classes are aspects that differ radically from the structure of textbooks. The computer-assisted instruction session presented to a student on a given day includes exercises from a selection of the strands appropriate to the student's grade level. In general the exercises are selected across the strands on a random basis according to a curriculum distribution, which is itself adjusted to be a convex mixture with a purely subjective distribution depending upon the individual student's strengths or weaknesses. The curriculum probability distribution is itself something not to be found in the general theory of curriculum or in textbooks, but is essential to the operational aspects of the kind of course I am describing. I do not here, however, go into the details of either the curriculum distribution or the individual student distribution, nor do I discuss the equally important problems of contingent tutoring when the student makes a mistake, or how a student is initially placed in the course.

The detailed use of learning theory occurs in monitoring and evaluating when

a student passes a mastery criterion for leaving a given equivalence class of exercises to move on to a more difficult concept or skill. Here we have been able to apply detailed ideas that go back to the framework of statistical learning theory introduced in Bill Estes' 1950 paper, but with an application he probably never originally had in mind: to the daily activity of students in schools in many parts of the country. Classical and simple ideas about mastery say that all that is needed is a test. On the basis of the relative frequency of correct answers in a given sample a decision is made as to whether the student shows mastery. But the first thing that one finds in the detailed analysis of data is that when a student is exposed to a class of exercises, even with previous instruction on the concepts involved, the student will show improvement in performance. The problem is to decide when the student has satisfied a reasonable criterion of mastery.

The emphasis here is on the data analysis of the actual learning, not on the setting of a normative, criterion of mastery. For this purpose some standard notation is introduced. Although in almost all of the exercises concerned, the variety of wrong student responses is large, I move at once to characterize responses abstractly as either correct or incorrect. With this restriction in mind I use the following familiar notation:

$A_{0,n}$ = event of incorrect response on trial n,

$A_{1,n}$ = event of correct response on trial n,

x_n = possible sequence of correct and incorrect responses from Trial 1 to n inclusive,

$q_n = P(A_{0,n})$, the mean probability of an error on Trial n,

$q = q_1$,

$q_{x,n} = P(A_{0,n}|x_{n-1})$.

Also, $\underline{A}_0$ and $\underline{A}_1$ are the corresponding random variables.

The learning model that Zanotti and I have applied to the situation described is one that very much fits into the family of models developed by Estes, Bush, Mosteller, and the rest of us working in the 1950s. In addition to q, there are two parameters to the model. One is a uniform learning parameter α that acts constantly on each trial, because the student is always told the correct answer; and the second is a parameter w, which assumes a special role when an error is made. This is one way to formulate the matter. A rather more interesting way perhaps is to put it in terms of the linear learning model arising from statistical learning theory analyzed very thoroughly in Estes and Suppes (1959, 1959a, 1959b). The parameter α corresonds to $1 - \theta$ in the linear model derived from statistical learning theory. The linear learning model derived from statistical learning theory puts all of the weight as such on the reinforcement. The generalization considered here is straightforward and can be found in earlier studies as well. In terms

of the two parameters α and w we may write the basic assumptions of the model in terms of the following two equations:

$$P(A_{0,n+1}|A_{0,n},x_{n-1}) = (1 - w)\alpha P(A_{0,n}|x_{n-1}) + \alpha w, \tag{1}$$

$$P(A_{0,n+1}|A_{1,n},x_{n-1}) = (1 - w)\alpha P(A_{0,n}|x_{n-1}). \tag{2}$$

It is then easy to prove by familiar methods that in terms of random variables:

$$E(\underline{A}_{0,n+1}) = \alpha^n q. \tag{3}$$

$$\text{Var}\,(\underline{A}_{0,n+1}) = \alpha^n q(1 - \alpha^n q). \tag{4}$$

Equation (3) just expresses in terms of random variables the familiar mean learning curve, which also holds in the learning model of statistical learning theory, but written in terms of θ rather than in terms of α. A variety of methods are available for estimating the three parameters of the model, namely, the parameters q, α, and w, but I do not go into such methods here.

In Figs. 1.1–1.4 I show four mean learning curves for third- and fourth-grade exercises in addition and multiplication of whole numbers and addition of fractions. The grade placement of each class of exercises is shown in the figure title,

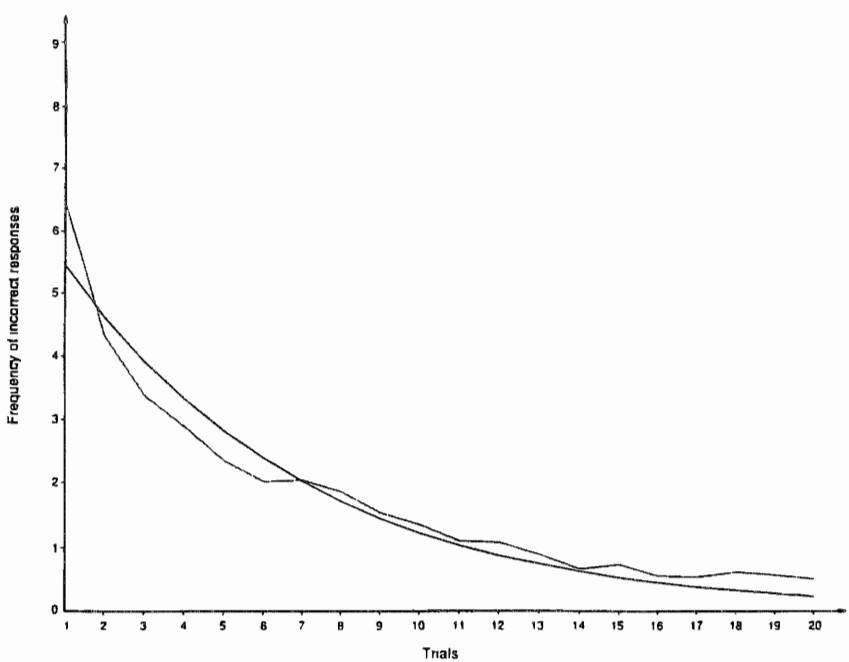

FIG. 1.1. Addition at grade-level 3.65, sample size = 612, $\hat{q}$ = 0.545, $\hat{\alpha}$ = 0.847.

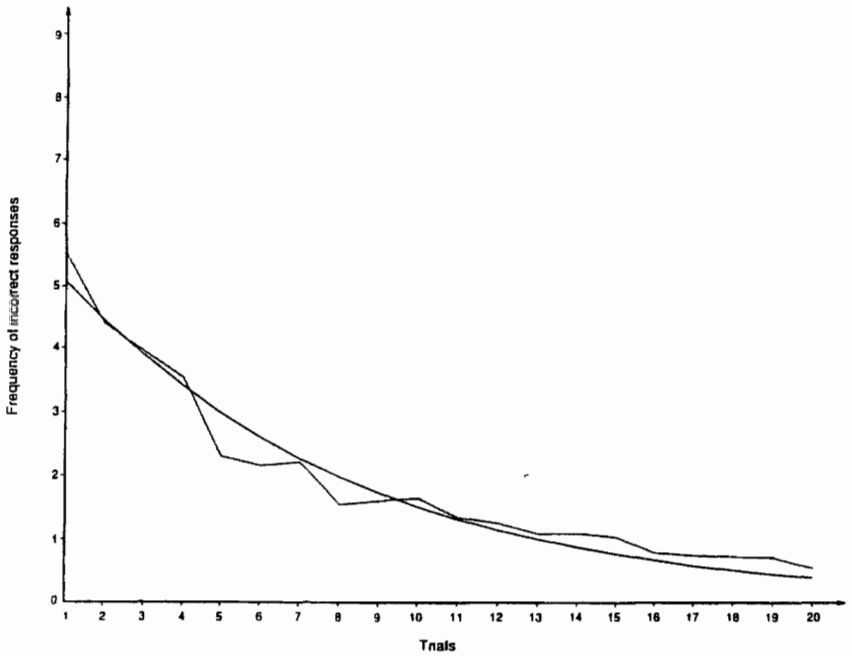

FIG. 1.2. Multiplication exercises at grade level 3.65, sample size = 487, $\hat{q} = 0.508$, $\hat{\alpha} = 0.875$.

for example, 3.65 for multiplication. But this does not mean that all the students doing these exercises were in the third grade, for with the individualization possible, students can be from several different chronological grade levels. The sample sizes on which the mean curves are based are all large, ranging from 406 to 616. The students do not come from one school and certainly are not in any well-defined experimental condition. On the other hand, all of the students were working in a computer laboratory run by a proctor in an elementary school so there was supervision of a general sort of the work by the students, especially in terms of schedule and general attention to task. In these mean learning curves and the sequential data presented later, the students who responded correctly to the first four exercises have been deleted from the sample size and from the data, because in terms of the mastery criterion used, students who did the first four exercises correctly in a given class were immediately moved to the next class in that strand. No further deletions in the data were made. For each figure the estimated initial probability q of an error and the estimated learning parameter α are given.

The data and theoretical curves shown in Figs. 1.1–1.4 represent four from a sample of several hundred, all of which show the same general characteristics;

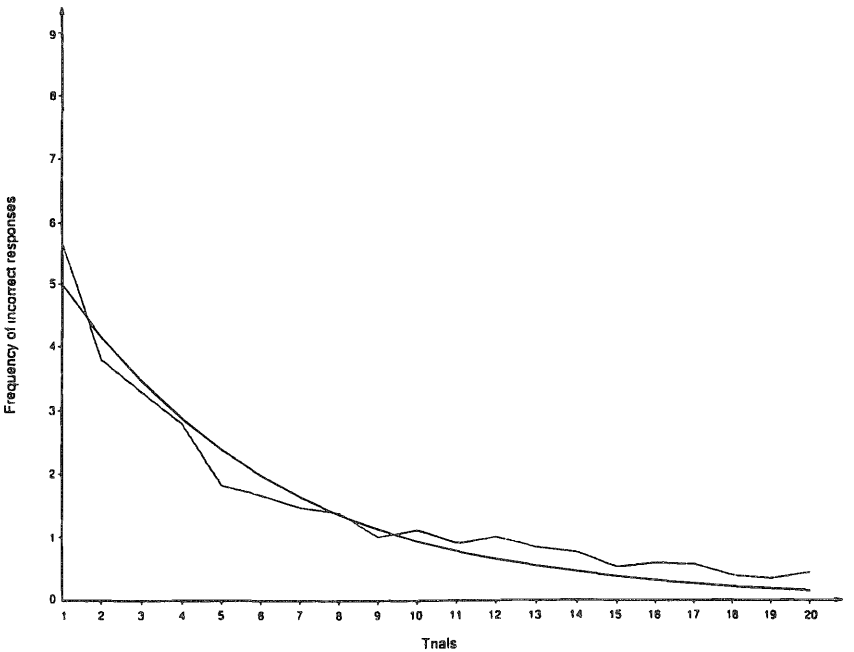

FIG. 1.3. Multiplication exercises at grade level 3.95, sample size = 406, $\hat{q} = .500$, $\hat{\alpha} = 0.831$.

namely, very rapid improvement in probability of a correct response as practice continues from the first trial onward. In most cases the student will have at least one intervening trial between exercises from a given class. So, for example, between two fraction exercises there might well intervene several different exercises, one a word problem, another a decimal problem, and so on. Also, it is probably true for all of the students that they had had some exposure by their classroom teacher to the concepts used in solving the exercises, but, as is quite familiar from decades of data on elementary-school mathematics, students show clear improvement in correctness of response with practice. In other words, learning continues long after formal instruction is first given. The most dramatic example of an improvement is in Fig. 1.4. This is not unexpected, because understanding and manipulation of fractions are among the most difficult concepts for elementary-school students to master in the standard curriculum.

In Tables 1.1–1.4, data from the same four classes of exercises are analyzed in terms of the more demanding requirement on the learning model to fit the sequential data. In the present case we looked at the first four responses with, as already indicated, the data for students with four correct responses deleted. This left a joint distribution of 15 cells, and in the case of the sequential data, the

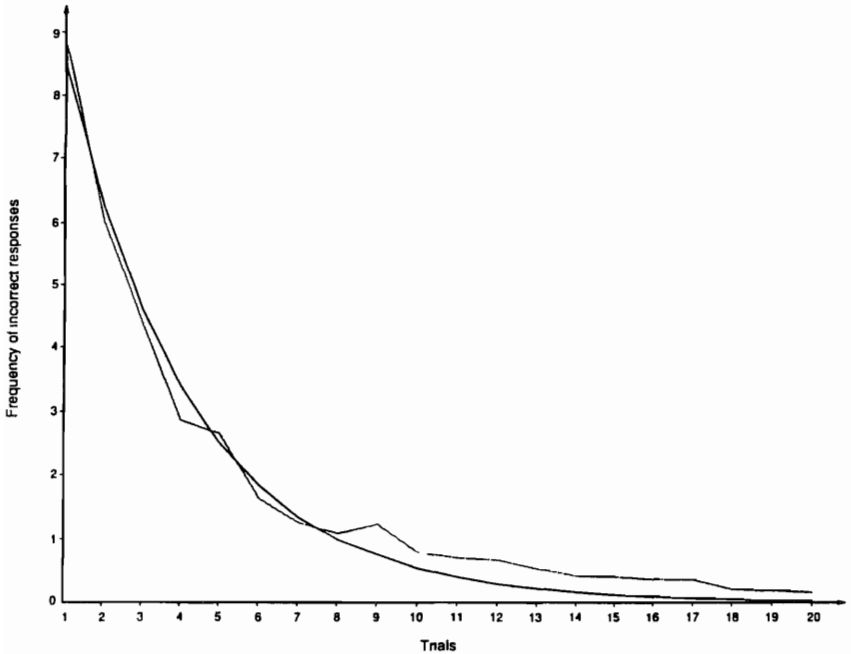

FIG. 1.4. Fraction exercises at grade level 4.40, sample size = 616, $\hat{q}$ = 0.849, $\hat{\alpha}$ = 0.736.

TABLE 1.1
Observed and Theoretically Expected Distribution for the First Four Responses to Addition Exercises at Level 3.65.
Sample Size - 612, DF = 11, $\hat{\omega}$ = 0.188, $\hat{q}$ = 0.406, $\hat{\alpha}$ = 0.750, $\chi 2$ = 14.39

Cell	*Observed*	*Expected*
0000	25	20.6
0001	36	35.0
0010	19	21.1
0011	75	70.6
0100	13	15.6
0101	34	39.3
0110	31	25.5
0111	163	151.6
1000	12	12.7
1001	19	27.3
1010	19	17.3
1011	60	80.0
1100	18	14.6
1101	49	48.3
1110	39	32.6

TABLE 1.2
Multiplication Exercises at Level 3.65.
Sample Size - 488, DF = 11, $\hat{\omega}$ = 0.125, $\hat{q}$ = 0.375, $\hat{\alpha}$ = 0.875, $\chi 2$ = 4.44

Cell	*Observed*	*Expected*
0000	17	17.9
0001	25	24.2
0010	20	18.9
0011	40	41.0
0100	19	16.1
0101	33	31.0
0110	30	25.1
0111	87	83.1
1000	13	14.7
1001	25	25.9
1010	16	20.8
1011	59	61.8
1100	15	18.7
1101	46	48.6
1110	43	40.2

theoretical computations involve the parameter w as well as α. For learning theorists the stringency of the test to fit such sequential data with only three parameters is well known. The data in each of the tables have 11 degrees of freedom because of the three parameters estimated from the data. The largest X^2 is for the addition exercises, but even here the X^2 is not significant at the 0.10 level. It is to my mind surprising that the data fit as well as they do.

It is to be noticed that the values of q and α are not the same for the mean learning curve and the joint distributions. This is because we fit each case

TABLE 1.3
Multiplication Exercises at Level 3.95.
Sample Size - 406, DF = 11, $\hat{\omega}$ = 0.188, $\hat{q}$ = 0.281, $\hat{\alpha}$ = 0.781, $\chi 2$ = 9.43

Cell	*Observed*	*Expected*
0000	12	9.4
0001	12	16.0
0010	9	10.5
0011	29	36.3
0100	10	8.6
0101	23	22.3
0110	21	15.6
0111	115	103.4
1000	8	8.2
1001	17	18.1
1010	13	12.4
1011	54	62.6
1100	8	11.5
1101	43	41.3
1110	32	29.7

TABLE 1.4
Fraction Exercises at Level 4.40.
Sample Size - 616, DF = 11, $\hat{\omega}$ = 0.313, $\hat{q}$ = 0.813, $\hat{\alpha}$ = 0.688, $\chi 2$ = 8.87

Cell	*Observed*	*Expected*
0000	76	73.5
0001	96	89.8
0010	34	38.8
0011	121	126.0
0100	21	21.7
0101	41	40.5
0110	18	21.1
0111	141	136.5
1000	6	7.7
1001	12	11.5
1010	9	5.5
1011	16	23.8
1100	5	4.2
1101	13	9.9
1110	7	5.5

directly and the best estimates were not precisely the same for the two different statistics, as is familiar in learning data. It is also apparent that if we used the same values for both mean learning curves and the sequential data, the fits would still be reasonably good in all four classes exhibited in Figs. 1.1–1.4 and Tables 1.1–1.4.

The main point of my presentation of these data is to show the viability of the learning models that originate in statistical learning theory and related work in the 1950s to nonexperimental school situations. I argue that the examples given show that much deeper application of mathematical and quantitative learning concepts and theories can be applied directly to school learning. This applicability of a very direct kind contrasts quite markedly to any attempt to apply in the same quantitative fashion learning ideas to be found in the earlier work of Hull or Skinner, for example, cited at the beginning of the previous section.

Robots that Learn

In this example of the application of statistical learning theory a more radical attitude is taken toward the use of the theory, because it is not an application directed at the analysis of data from human or animal performance, but rather uses the theory as the basis of a built-in mechanism to smooth out and make robust the performance of a robot that is learning a new task. The idea of such instructable robots, on which I have now worked for a number of years with Colleen Crangle and others (Crangle & Suppes, 1987, 1990; Maas & Suppes, 1985; Suppes & Crangle, 1988) is organized around two leading ideas. First, robots should be able to learn from instructions given in ordinary, relatively

vague English, just as we expect children and apprentices to do so in learning a new task. Second, quite apart from the vagueness of the English, even the conceptualization of the task is often relatively vague and qualitative in character. The use of precise coordinates is not easily available to the instructor and consequently probability distributions provide a method of robustly adjusting to the nature of the task—probability distributions, I should emphasize, that are dynamically changing as a consequence of learning.

The approach we use to these matters follows directly from work of my own on learning models for a continuum (Suppes, 1959a, 1959b), which very much falls within the framework of statistical learning theory. The leading idea is that on any trial the robot has a smoothing distribution $\kappa_{\mu,\sigma}(x)$, which earlier I called a smearing distribution. (As should be evident, μ is the mean of the distribution and σ^2 its variance.) The idea is that this is a distribution that is smeared around the conceptually ideal response. The probability of a response lying between a and b is $\int_a^b \kappa_{\mu,\sigma}(x)dx$. Notice that I am implicitly assuming, and now do so explicitly, that the applications are confined to one dimension. In fact, for the present discussion I restrict myself to the beta distribution on the open interval (0,1)

$$\beta(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}$$

where, as is familiar, the distribution requires that $\alpha,\beta > 0$. Moreover the mean μ and the variance σ^2 are:

$$\mu = \frac{\alpha}{\alpha + \beta}, \qquad \sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

Note that the uniform distribution is a special case of the beta distribution when $\alpha = \beta = 1$. In our present work Crangle and I have restricted ourselves to one dimension although there is considerable interest in the two-dimensional applications. But the learning–theoretic ideas become more complicated rather rapidly. A typical one-dimensional task that we may think of for purposes of illustration is a robot opening a door. Its problem is to position itself along an interval parallel to the door when closed. Within that framework we think of three kinds of feedback or reinforcement to the robot: positional feedback, accuracy feedback, and congratulatory feedback, where each of these different kinds of feedback are expressed in ordinary qualitative language. In the case of positional feedback, we have three range constants indicating how much the means should be moved. The constants are qualitatively thought of as large, medium, and small, with the large parameter being 2, the medium 1, and the small 0.5. (The exact value of these parameters is not important.) In this framework we have then the relations between feedback and learning shown in Table 1.5.

Notice that all of the positional feedback commands are not given. It is understood that for each of those to the left, there is a corresponding one to the

TABLE 1.5
Relation Between Feedback and Learning

Feedback/Reinforcement	*Learning*
Much further left!	$\mu_{n+1} = \mu_n - 2\sigma$
Right just a little!	$\mu_{n+1} = \mu_n + 0.5\sigma$
Move to the left!	$\mu_{n+1} = \mu_n - \sigma$
Be more careful!	$\mu_{n+1} = \sigma_n^2$
No need to be so cautious!	$\mu_{n+1} = \sqrt{\sigma_n}$
Just right!	$\mu_{n+1} = r_n$ $\sigma_{n+1}^2 = \sigma_n^2$

right, and vice versa. In the case of the positional feedback the variance stays the same, and only the mean is changed. In the case of accuracy feedback, shown in the second part of Table 1.5, only the variance is changed, not the mean. The algebraic expression of the change in variance reflects the fact that we are restricting ourselves in the present example to the open interval (0, 1). Finally, congratulatory feedback is to indicate that the distribution being used is satisfactory. Its location and accuracy is about right and so the mean should be the last response, that is $\mu_{n+1} = r_n$, and the variance stays the same. In order to reduce the number of figures, in Fig. 1.5 I show at each step only the distribution on the basis of which the response is made. The response is shown by a heavy black vertical line and then the verbal feedback that follows use of this distribution to the right of the graph. The graphs are to be read in sequence starting at the top and moving to the right from the upper left-hand corner. I should note that the initial instruction not shown here is "Go to the door," which is also repeated at the beginning of each new trial.

As the graphs are meant to show, with appropriate reinforcement and further trials, the robot ends up with a reasonable distribution for the task of opening the door. It might seem that we should not go to all this trouble of repeated trials and the use of English, when one could just program in from the beginning the correct distribution. This has been very much the attitude in much of the work in computer science on robots, but a little reflection on the wide world of practical activities will indicate that it is a very limited approach to performing most tasks. Complicated problems requiring sophisticated motor and perceptual skills are really never handled by the explicit introduction of precise parameters. Apprentices learn from watching and getting instruction of a general sort from a master craftsman. A youngster learning to play the piano or learning tennis is not told

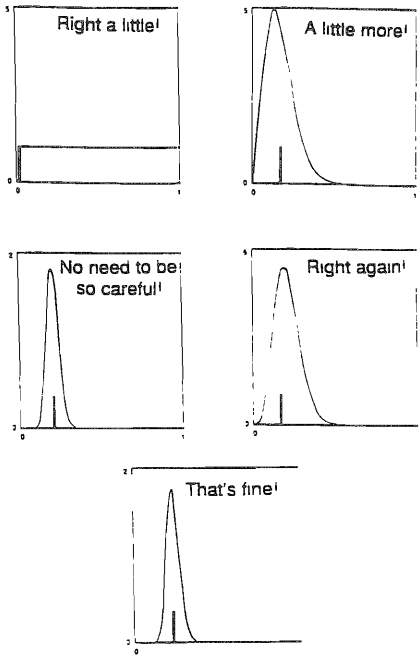

FIG. 1.5. Illustrations of the changes in probability distribution by an instructable robot learning to open a door.

how to set parameters but is given practice and repeated qualitative instruction on ways to improve the performance. I am not suggesting by these remarks that we have reached a very sophisticated level in the instruction of robots, but just that the road that lies ahead for the kind of work outlined here seems to be conceptually the right approach to a great many tasks we will expect robots to perform in the future. The central point I want to make is that the way we have thought about learning in the past seems to me very appropriate for thinking about learning in the future as far as robots are concerned. In this simple example, for instance, I have certainly bypassed all of the problems of laying out the exact cognitive and perceptual capabilities of the robots to whom the theory is supposed to apply. When one turns to actual examples, it is clear that the cognitive and perceptual capabilities will be severely limited but can in many cases be adequate to specific tasks. It is my own conjecture that the kind of ideas initiated by Bill Estes and others in the 1950s will have a useful, and in many cases central, role to play in the training of robots for specific tasks, not merely in the distant future but rather soon.

Learning Boolean Functions

In this third example I turn to purely theoretical issues, but ones that have been the focus of a number of papers in computer science in the last few years, especially by Valiant (1984, 1985). Much of the work of Valiant and others—I mention especially the recent long article by Blumer, Ehrenfeucht, Haussler, and Warmuth (1989)—has been on learnability of various classes of concepts, including especially Boolean classes. The thrust of this theoretical work has essentially been to show under what conditions concepts can be learned in polynomial time or space. These general polynomial results are important, but I would like to focus here on reporting joint work with Shuzo Takahashi in which we are much more concerned with detailed results of the kind characteristic of learning theory as applied to human behavior. Secondly, the work of Valiant has emphasized the importance of considering only restricted Boolean functions that have some special syntactic form, for example, disjunctive normal form. Although I state some general results that Takahashi and I have obtained, I mainly concentrate on examples, in order to show the power of imposing a learning hierarchy whenever pertinent conditions obtain.

For the analysis of learning Boolean functions, it is very natural to go back for the learning process to one of Bill Estes' most important papers, the 1959 one entitled "Component and pattern models with Markovian interpretations." I especially have in mind Bill's introduction of the pattern model in this paper. The assumptions about the pattern model close to his that we use are these: First, one stimulus pattern is presented on each trial; second, initially all patterns are unconditioned; third, conditioning occurs with Probability 1 to the correct response on each trial. This last assumption is a specialization of the general formulation Estes uses in which conditioning occurs with Probability c. In this application, intended for machine learning, it is appropriate to set $c = 1$. Fourth, there are two responses, 1 for the pattern being an instance of the concept, and 0 for its not so being. Finally, for the initial discussion I make a sampling assumption that is much more special than any assumed in Bill's 1959 paper, but one that is useful for the purposes of clarifying the role of a hierarchy in the learning of Boolean functions. This is the strong assumption that there is a complete sampling of patterns without replacement.

The intuitive idea of the hierarchy is easily introduced, and the reasons for it as well. Suppose we are interested in learning an arbitrary Boolean function of n variables. Then there are 2^n patterns. Without introducing a hierarchy there is, in the general case, no natural way to avoid sampling all of the patterns if we want to learn the concept completely. Even when we want to learn the concept incompletely, but with reasonable confidence of a correct response, if the patterns are distributed more or less uniformly, then learning by sampling each pattern is not feasible—not feasible in the technical sense that it is exponential in n.

On the special learning assumptions introduced, and I emphasize particularly the assumption that patterns are sampled without replacement, it is easy to prove some general results. As a matter of notation let

N_t = number of trials for nonhierarchical learning of term t,
H_t = number of trials for hierarchical learning of t,
V_t = number of distinct variables in t,
S_t = set of function symbols in t,
$A(f)$ = number of arguments in function t.

We may then easily show that

$$N_t = 2^{V_t}$$
$$H_t = \sum_{f \in S_t} 2^{A(f)}.$$

Here are some simple examples:

For $t = f(x,y), N_t = 2^2 = 4$ and $H_t = 2^2 = 4$.
For $t = f(g(x_1,x_2),h(x_1,x_2,x_3))$, $N_t = 2^3 = 8$ and $H_t = 2^2 + 2^2 + 2^3 = 16$.
But for $t = f(g(x_1,x_2,x_4), g(x_2,x_3,x_4))$, $N_t = 2^4 = 16$ and $H_t = 2^2 + 2^3 = 12$.

Note that in each of these examples, f, g, and h are arbitrary Boolean functions. The exact form of the function does not matter. For instance, in the first example f could be Boolean intersection, Boolean disjunction, or Boolean exclusive or. Here is the best sort of hierarchical example:

$$t = h(g(f(x_1, x_2), f(x_3, x_4)), g(f(x_5, x_6), f(x_7, x_8)))$$
$$N_t = 2^8 = 256$$
$$H_t = 3 \times 2^2 = 12$$
$$= 4 \log_2 V_t$$
$$= 4 \log_2 8 = 4 \times 3 = 12.$$

On the other hand, here is the worst sort of hierarchical example:

$$t = h(g(f_1(x_1, x_2), f_2(x_1, x_2)), g_2(f_3(x_1, x_2), f_4(x_1, x_2)))$$
$$N_t = 2^2 = 4$$
$$H_t = 7 \times 2^2 = 28.$$

Introducing one more matter of notation, let $F_t = |S_t|$ be the number of function symbols in t. We may then state the following theorems that are easily proved.

Theorem 1. *For terms with only binary functions, if* $V_1 \geq 4$ *and* $V_t \geq F_t$ *then* $H_t \leq N_t$.

Notice that Theorem 1 establishes a general inequality for hierarchical versus nonhierarchical formulations with restriction to binary functions. Terms so restricted we call *binary terms.*

Theorem 2. *For binary terms, if* $V_t \geq 4$ *and* $F_t/V_t \leq r$ *then* $H_t/N_t \leq r$.

More than a decade ago (Suppes, 1977) I was emphasizing the importance of hierarchies for efficiency in learning, but at that time I did not appreciate how great the gain might be. I end by describing one simple example that is quite dramatic. Let the number of individual variables be 1024—obviously I pick this number to have a nice power of 2 but the exact number is of no importance. Then the number of patterns is 2^{1024}, which on the basis of individual sampling would take more time to complete than the universe has yet experienced. The number of trials, although finite, is unthinkably large. However, I now introduce the following three assumptions. First, let the hierarchy be of the form of the best example given above but such that each occurrence of the binary function f is now a distinct function f_i. Second, let there be learner control of stimulus sampling, so that the learner can sample exactly the pattern desired; and third, let the processing be parallel. Then the trials to learn an arbitrary Boolean function expressible in this particular hierarchy by binary functions is just 40. This is a reduction from 2^{1024} of more than 300 orders of magnitude! More generally, for n the number of variables, the trials to learn for this best hierarchical example with parallel processing is $4 \log_2 n$.

The proof in the general case is quite simple. The four patterns for parallel processing at the initial stage and recursively then for each next level of the hierarchy can be shown as follows:

x_1x_2	x_3x_4	.....	$x_{n-1}x_n$
11	11	.....	11
10	10	.....	10
01	01	.....	01
00	00	.....	00

Second, it is easy to show that the number of learning levels in the hierarchy is simply $\log_2 n$. This last example with $n = 1024$ is admittedly a very special case, but, even in a very special case, to obtain a reduction of 300 orders of magnitude in the number of learning trials is something not easy to find in any domain of knowledge under any circumstances whatsoever. We are all intuitively convinced of the importance of hierarchies in the organization of large-scale learning in the schools, but it is very desirable to have a dramatic artificial example to show their power.

THE FUTURE

It is obvious enough that I think statistical learning theory will continue to flourish in the future and have a place of importance. Over the past generation we

have raised a lot of cognitive psychologists who know very little about learning of a serious sort, but the new generation is properly learning all about connectionism and neural nets. A natural question is whether the pattern model, as just used in the discussion of learning Boolean functions, will be replaced by neural nets. There is an important distinction to be made here—a distinction that is critical in understanding why statistical learning theory will continue to have a significant future. This distinction is that between methods of implementing learning, and analysis in terms of best results possible. In describing the pattern model I have not at all tried to describe how it would be implemented in hardware or software of either the biological or electronic kind. It is easy enough to show, and should be apparent, that the learning results for Boolean functions just discussed cannot be improved by any new network configuration. In fact, it would undoubtedly be the case that under many kinds of network analysis the learning rate would not be so rapid as that shown for the pattern model. Granted, I mean the pattern model with the strong assumption that the conditioning parameter $c = 1$. This example of course was meant to be artificial, but the same kind of very strong hierarchical results would obtain under a broad range of sampling assumptions about the patterns. What is said there about parallel processing of patterns is critical to efficient learning of a biological as well as electronic kind. The most obvious example is the human vision system, although I am not pretending that a simple analysis in terms of Boolean functions will suffice to provide a detailed theory.

Parallel processing, it seems to me, is the important concept to add to Estes' classical pattern model. This addition will represent an important modification for the future. In contrast, the use of neural networks is not an extension, or in contradiction to, the pattern model, but represents a lower level of abstraction, what I would call a schematic level of implementation. I hasten to add that I strongly favor this lower level of abstraction, for a full account of learning needs a theory of the biological or electronic processing of both external stimuli and internal computations.

On the other hand, as much as such an extended theory is desirable, we can easily be misled into understanding the difficulty of developing it or the usefulness of simplifying abstractions like the pattern model. A pluralism of levels of abstraction and of corresponding models is, in my judgment, a permanent feature of any science of complex phenomena. It is naive and mistaken to think we shall find the one true complete theory of learning based on accurate details of how neurons and their connections actually work. Many different levels of theorizing will continue to be of value in many different domains. There is every reason to think that the kinds of applications of statistical learning theory described in the previous section will have a robust future.

More generally, the classical concepts of association, generalization, and discrimination will be extended, but it seems likely that these basic concepts will continue to play the role in the psychology of learning that the concepts of

classical mechanics such as force, mass, and acceleration have played for 200 years of physics. It is not that physics has not developed many new concepts and theories, it is, rather, that once fundamental concepts are put in some reasonable mathematical framework and are recognized as having great generality, they do not disappear. Such will be the future of the fundamental ideas of statistical learning theory.

ACKNOWLEDGMENTS

I am grateful to Duncan Luce for a number of helpful comments on an earlier draft.

REFERENCES

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery, 36,* 929–965.

Bush, R. R., & Mosteller, F. (1951a). A mathematical model for simple learning. *Psychological Review, 58,* 313–323.

Bush, R. R., & Mosteller, F. (1951b). A model for stimulus generalization and discrimination. *Psychological Review, 58,* 413–423.

Crangle, C., & Suppes, P. (1987). Context-fixing semantics for instructable robots. *International Journal of Man-Machine Studies, 27,* 371–400.

Crangle, C., & Suppes, P. (1990). Introduction dialogues: Teaching new skills to a robot. *Proceedings of the NASA Conference on Space Telerobotics, January 31–February 2, 1989,* pp. 91–101.

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

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., & Burke, C. J. (1953). A theory of stimulus variability in learning. *Psychological Review, 60,* 276–286.

Hilgard, E. R., & Marquis, D. G. (1940). *Conditioning and Learning.* New York: Appleton-Century

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

Maas, R., & Suppes, P. (1985). Natural-language interface for an instructable robot. *International Journal of Man-Machine Studies, 22,* 215–240.

Miller, G. A. (1952). Finite Markov processes in psychology. *Psychological Review, 17,* 149–167

Miller, G. A., & McGill, W. J. (1952). A statistical description of verbal learning. *Psychometrika, 17,* 369–396.

Skinner, B. F. (1938). *The behavior of organisms.* New York: Appleton-Century.

Suppes, P. (1959a). A linear model for a continuum of responses. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory* (pp. 400–414). Stanford, CA: Stanford University Press.

Suppes, P. (1959b). Stimulus sampling theory for a continuum of responses. In K. J. Arrow, S. Karlin, & P. Suppes (Eds.), *Mathematical methods in the social sciences* (pp. 348–365). Stanford, CA: Stanford University Press.

Suppes, P. (1977). Learning theory for probabilistic automata and register machines, with applications to educational research. In H. Spada & W. F. Kempf (Eds.), *Structural models of thinking and learning, Proceedings of the 7th ISPN-Symposium on Formalized Theories of Thinking and Learning and their Implications for Science Instruction* (pp. 57–79). Bern, Switzerland: Hans Huber.

Suppes, P. (1989). Current directions in mathematical learning theory. In E. E. Roskam (Ed.), *Mathematical psychology in progress* (pp. 3–28). Berlin: Springer-Verlag.

Suppes, P., & Crangle, C. (1988). Context-fixing semantics for the language of action. In J. Dancy, J. M. E. Moravcsik, & C. C. W. Taylor (Eds.), *Human agency: Language, duty and value* (pp. 47–76, 288–290). Stanford, CA: Stanford University Press.

Valiant, L. G. (1984). A theory of the learnable. *Communic. ACM, 27,* 1134–1142.

Valiant, L. G. (1985). Learning disjunctions of conjunctions. In *Proceedings of the 9th International Conference on Artificial Intelligence,* vol 1 (pp. 560–566). San Mateo, CA: Morgan Kaufmann.

Wiener, N. (1948). *Cybernetics.* New York: Wiley.