CHAPTER 3

Mechanisms of Skill Acquisition and the Law of Practice

A. Newell and P. S. Rosenbloom, Carnegie Mellon University

INTRODUCTION[1]

Practice makes perfect. Correcting the overstatement of a maxim: Almost always, practice brings improvement, and more practice brings more improvement. We all expect improvement with practice to be ubiquitous, though obviously limits exist both in scope and extent. Take only the experimental laboratory: We do not expect people to perform an experimental task correctly without at least some practice; and we design all our psychology experiments with one eye to the confounding influence of practice effects.

Practice used to be a basic topic. For instance, the first edition of Woodworth (1938) has a chapter entitled "Practice and Skill." But, as Woodworth [p. 156] says, "There is no essential difference between practice and learning except that the practice experiment takes longer." Thus, practice has not remained a topic by itself but has become simply a variant term for talking about learning skills through the repetition of their performance.

With the ascendence of verbal learning as the paradigm case of learning, and its transformation into the acquisition of knowledge in long-term memory, the study of skills took up a less central position in the basic study of human behavior. It did not remain entirely absent, of course. A good exemplar of its

[1]This chapter relies on the data of many other investigators. We are deeply grateful to those who made available original data: John Anderson, Stu Card, Paul Kolers, Tom Moran, David Neves, Patrick Rabbitt, and Robert Seibel. We are also grateful to John Anderson, Stu Card, Clayton Lewis, and Tom Moran for discussions on the fundamental issues and, especially, to Clayton Lewis for letting us read his paper, which helped to energize us to this effort.

continued presence can be seen in the work of Neisser, taking first the results in the mid-sixties on detecting the presence of ten targets as quickly as one in a visual display (Neisser, Novick, & Lazar, 1963), which requires extensive practice to occur; and then the recent work (Spelke, Hirst, & Neisser, 1976) showing that reading aloud and shadowing prose could be accomplished simultaneously, again after much practice. In these studies, practice plays an essential but supporting role; center stage is held by issues of preattentive processes, in the earlier work, and the possibility of doing multiple complex tasks simultaneously, in the latter.

Recently, especially with the articles by Shiffrin & Schneider (1977; Schneider & Shiffrin, 1977), but starting earlier (LaBerge, 1974; Posner & Snyder, 1975), emphasis on *automatic* processing has grown substantially from its level in the sixties. It now promises to take a prominent place in cognitive psychology. The development of automatic processing seems always to be tied to extended practice and so the notions of skill and practice are again becoming central.

There exists a ubiquitous quantitative law of practice: It appears to follow a power law; that is, plotting the logarithm of the time to perform a task against the logarithm of the trial number always yields a straight line, more or less. We shall refer to this law variously as the *log-log linear learning law* or the *power law of practice*.

This empirical law has been known for a long time; it apparently showed up first in Snoddy's (1926) study of mirror-tracing of visual mazes (see also Fitts, 1964), though it has been rediscovered independently on occasion (DeJong, 1957). Its ubiquity is widely recognized; for instance, it occupies a major position in books on human performance (Fitts & Posner, 1967; Welford, 1968). Despite this, it has captured little attention, especially theoretical attention, in basic cognitive or experimental psychology, though it is sometimes used as the form for displaying data (Kolers, 1975; Reisberg, Baron, & Kemler, 1980). Only a single model, that of Crossman (1959), appears to have been put forward to explain it.[2] It is hardly mentioned as an interesting or important regularity in any of the modern cognitive psychology texts (Calfee, 1975; Crowder, 1976; Kintsch, 1977; Lindsay & Norman, 1977). Likewise, it is not a part of the long history of work on the *learning curve* (Guilliksen, 1934; Restle & Greeno, 1970; Thurstone, 1919), which considers only exponential, hyperbolic, and logistic functions. Indeed, a recent extensive paper on the learning curve (Mazur & Hastie, 1978) simply dismisses the log-log form as unworthy of consideration and clearly dominated by the other forms.

[2]But see Suppes, Fletcher, and Zanotti (1976), who do develop a model yielding a power law for instructional learning, though their effort appears independent of a concern with the general regularity. Unfortunately, their description is too fragmentary and faulty to permit it to be considered further.

The aim of this chapter is to investigate this law. How widespread is its occurrence? What could it signify? What theories might explain it? Our motivation for this investigation is threefold. First, an interest in applying modern cognitive psychology to user-computer interaction (Card, Moran, & Newell, 1980a; Robertson, McCracken, & Newell, in press) led us to the literature on human performance, where this law was prominently displayed. Its general quantitative form marked it as interesting, an interest only heightened by the apparent general neglect of the law in modern cognitive psychology. Second, a theoretical interest in the nature of the architecture for human cognition (Newell, 1980) has led us to search for experimental facts that might yield some useful constraints. A general regularity such as the log-log law might say something interesting about the basic mechanisms of turning knowledge into action. Third, an incomplete manuscript by Clayton Lewis (no date) took up this same problem; this served to convince us that an attack on the problem would be useful. Thus, we welcomed the excuse of this conference to take a deeper look at this law and what might lay behind it.

In the next section we provide many examples of the log-log law and characterize its universality. In the following section we perform some basic finger exercises about the nature of power laws. Then we investigate questions of curve fitting. In the next section we address the possible types of explanations for the law; and we develop one approach, which we call the *chunking theory of learning*. In the final section, we sum up our results.

THE UBIQUITOUS LAW OF PRACTICE

We have two objectives for this section. First, we simply wish to show enough examples of the regularity to lend conviction of its empirical reality. Second, the law is generally viewed as associated with *skill*, in particular, with perceptual-motor skills. We wish to replace this with a view that the law holds for practice learning of all kinds. In this section we present data. We leave to the next section issues about alternative ways to describe the regularity and to yet subsequent sections ways to explain the regularity.

We organize the presentation of the data by the subsystem that seems to be engaged in the task. In Table 1.1 we tabulate several parameters of each of the curves. Their definitions are given at the points in the chapter where the parameters are first used.

Perceptual-Motor Skills

Let us start with the historical case of Snoddy (1926). As remarked earlier, the task was mirror-tracing, a skill that involves intimate and continuous coordination of the motor and perceptual systems. Figure 1.1 plots the log of performance on the vertical axis against the log of the trial number for a single subject.

TABLE 1.1
Power Law Parameters for the (Log-Log) Linear Data Segments

Data Set	Power Law $T = BN^{-\alpha}$		
	B	α	r^2
Snoddy (1926)	79.20	.26	.981
Crossman (1959)	170.1	.21	.979
Kolers (1975) - Subject HA	14.85	.44	.931
Neisser et al. (1963)			
Ten targets	1.61	.81	.973
One target	.68	.51	.944
Card, English & Burr (1978)			
Stepping keys - Subj. 14	4.95	.08	.335
Mouse - Subj. 14	3.02	.13	.398
Seibel (1963) - Subject JK	12.33	.32	.991
Anderson (1980) - Fan 1	2.358	.19	.927
Moran (1980)			
Total time	30.27	.08	.839
Method time	19.59	.06	.882
Neves & Anderson (in press)			
Total time - Subject D	991.2	.51	.780
The Game of Stair			
Won games	1763	.21	.849
Lost games	980	.18	.842
Hirsch (1952)	10.01	.32	.932

The first important point is:

- The law holds for performance measured as the *time* to achieve a fixed task.

Analyses of learning and practice are free a priori to use any index of performance (e.g., errors or performance time, which decrease with practice; or amount or quality attained, which increase with practice). However, we focus exclusively on measures of performance time, with quality measures (errors, amount, judged quality) taken to be essentially constant. Given that humans can often engage in tradeoffs between speed and accuracy, speed curves are not definable without a specification of accuracy, implicit or otherwise.[3] As we illustrate later, the log-log law also appears to hold for learning curves defined

[3]Snoddy used an indicator, 1/(time+errors), and we have replotted the figure using time+errors. This strikes the modern eye as incongruous, adding together apples and oranges. In fact, the measure is almost purely performance time. Snoddy was endeavoring to cope with the speed/accuracy tradeoff. He fixed the error rate to be equal to the performance time (in seconds) and had the subject work faster or slower in order to hold the error rate at that level. Thus the error rate bore a fixed average relationship to time; and adding the actual value of the errors to the performance time was a way of compensating for momentary shifts in the speed/accuracy tradeoff.

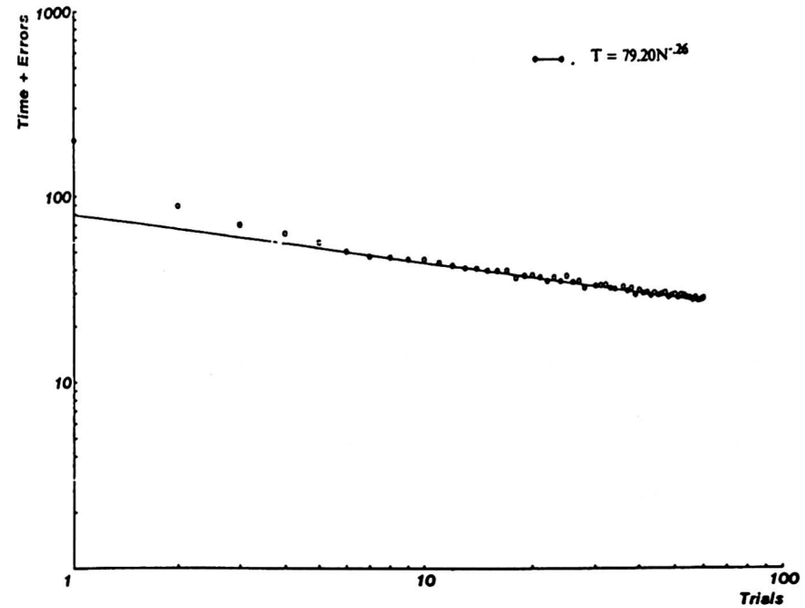

FIG. 1.1. Learning in a mirror tracing task (log-log coordinates). Replotted from Snoddy (1926).

on other performance criteria. Though significant for understanding the cause of the power law, we only note the existence of these other curves.

Several other things can be noted in Fig. 1.1, which show up generally in the other curves.

- The points are sparse at the left and become denser to the right. This arises from taking the log of the trial number. Even when trials are aggregated into blocks, this is usually done uniformly in linear space. Thus, this is just an artifact of the display.
- There is systematic deviation at one end. Here it is the beginning. Snoddy made a lot of this initial deviation, though we need not follow him in this. As we shall see, systematic deviation can occur at either end.
- There is little doubt that the bulk of the curve lies along a line in log-log space. This arises in part because of the relatively large number of points available.[4] The curves are for an individual, not for grouped data. This is not a condition of the law, but it shows that the law holds for individual data.

[4]Obvious deviations at the ends of the empirical curves were eliminated before the fits in Table 1.1 were computed. The equations therefore primarily represent this linear portion of the curve. The solid line in Fig. 1.1 (and in the following figures) reflects this fit.

• Data are rarely presented on many subjects, though in some cases such data exists and (apparently) is robust. For instance, Snoddy took his curve as diagnostic and appears to have gathered it on large numbers of individuals, though he never reported any mass of data.

In Table 1.1 we tabulate several critical features of the Snoddy data. The following equations describe the power law in linear and log-log spaces:

$$T = BN^{-\alpha} \tag{1}$$

$$\log(T) = \log(B) - \alpha \log(N) \tag{2}$$

B is the performance time on the first trial ($N = 1$) and α is the slope of the line (i.e., the learning rate). A positive value of α (e.g., .26 for the curve of Fig. 1.1) indicates a decreasing curve, because we have located the minus sign in the equation itself.

Another example from a task that appears to involve intimate motor-perceptual coordination is shown in Fig. 1.2. This is Crossman's (1959) famous data on the manufacture of cigars by female operators using a cigar-making machine. Noteworthy is the number of trials, namely, up to 20 million cigars.

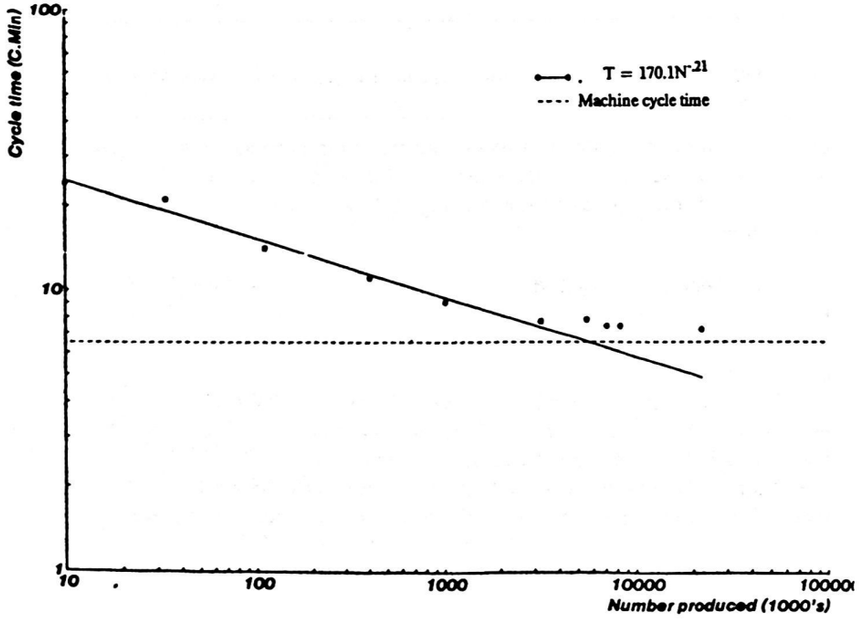

FIG. 1.2. Cross-sectional study of learning in cigar manufacturing (log-log coordinates). Replotted from Crossman (1959).

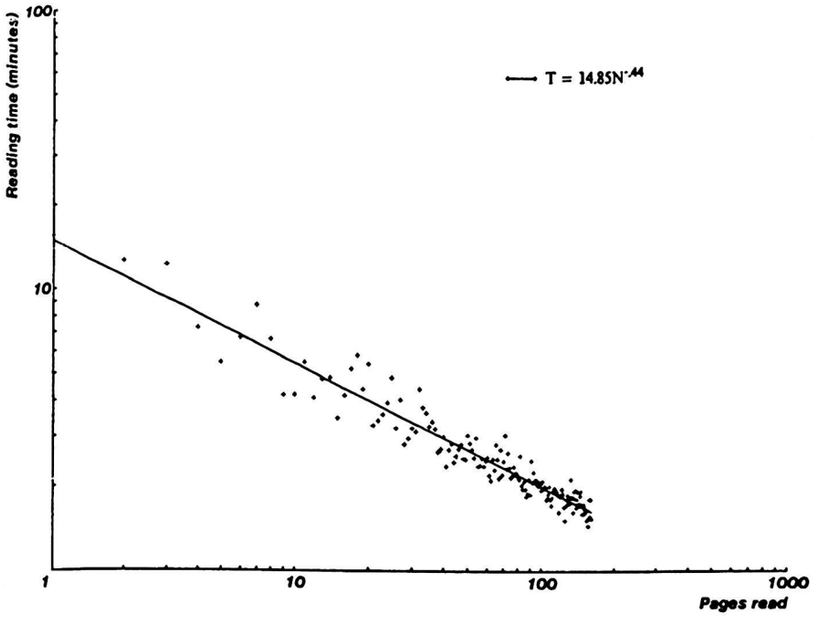

FIG. 1.3. Learning to read inverted text (log-log coordinates). Plotted from the original data for Subject HA (Kolers, 1975).

Also, there is a known lower bound for the performance time, namely, the cycle time of the machine. The curve eventually deviates from the log-log line, flattening out in submission to physical necessity. Still, practice followed the law for almost 3 million trials (and 2 years). Furthermore, additional small improvements continued; and it would be foolish indeed to predict that no further improvements would occur. Crossman's data differs from all other data in being cross-sectional (i.e., different individuals make up each point).

Perception

Figure 1.3 shows the data from one subject (of eight) in Kolers's well known studies on reading graphically transformed text (Kolers, 1975). Here, the transformation is inversion of each line around its horizontal axis. The task of the subject is to read many pages of such text for comprehension. Reading in general is a complex task, but the difficulties here are clearly strongly perceptual, being caused primarily by the perceptual transformation. Without inversion, reading is much faster and improves hardly at all (though we don't show Kolers's control data on this). In any event, as the figure shows, learning is log-log linear.

Figure 1.4 shows some data replotted from a paper by Neisser, Novick, & Lazar (1963). The task consisted of finding any of multiple targets in pages of

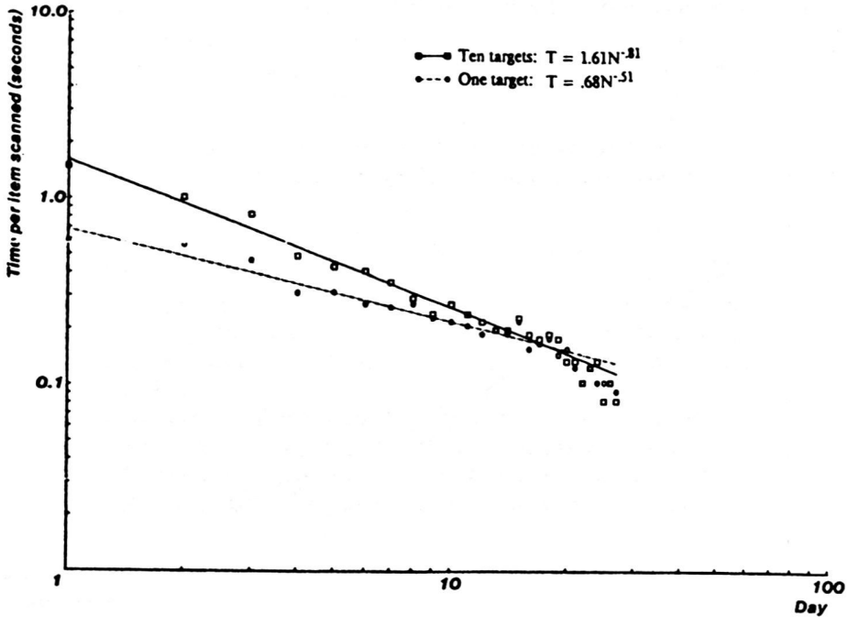

FIG. 1.4. Learning to scan for visual targets (log-log coordinates). Replotted from Neisser, Novick, & Lazar (1963).

letters. The result was that, with practice, identification time becomes essentially independent of the size of the target set. As Fig. 1.4 shows, this data also follows the log-log law, though there seems to be a slight drop at the end. These two curves (scanning for one target and for ten targets) represent the two bounding conditions of the five used in the experiment. Each curve is the average of six subjects. One of the reasons for exhibiting these particular curves is to point out that much learning data in the literature fits the log-log law, even though it has not been plotted that way.

Motor Behavior

Figure 1.5 is from a task where a subject sees a target mark appear on a video terminal and has to position the cursor at that mark (Card, English, & Burr, 1978). Four different pointing devices were used: a mouse, which permits a smooth pointing motion isomorphic to the motion of the cursor; a joystick; a set of stepping keys; and a set of text keys, which allow movement by paragraph, word, etc. Some of these devices are well described by Fitts's law (Fitts, 1954); some have a different structure. The two curves in Fig. 1.5 show the mouse and stepping key data for one subject, averaged over blocks of 20 trials (excluding

errors). For all of the devices, the total performance time follows the law, though the degree of variability increases as one moves from the Fitts's law devices (the mouse) toward the other ones.

Elementary Decisions

Figure 1.6 is from a task designed by Seibel (1963) to probe the dependence of reaction time on the number of alternatives. It followed in the wake of the work by Hick (1952), Hyman (1953), and others showing that choice RT was linear in the information (bits) required to select the response, at least for small ensembles (up to 3 or 4 bits). The subject's 10 fingers rested on 10 response keys (shaped to fit the natural position of the resting hand) and looked at 10 stimulus lights that were configured isomorphically to the keys. A subset of the lights would turn on, and the subject was to strike the corresponding keys. There are 1023 ($2^{10} - 1$) different subsets of the lights; hence, the arrangement achieves a choice RT task of 10 bits. For our purposes what is interesting is that the learning over a large number of trials (40,000) was log-log linear, though at the end the curve flattens out. This is data for a single subject, averaged over blocks of 1023 trials; approximately the same behavior was shown by each of three subjects.

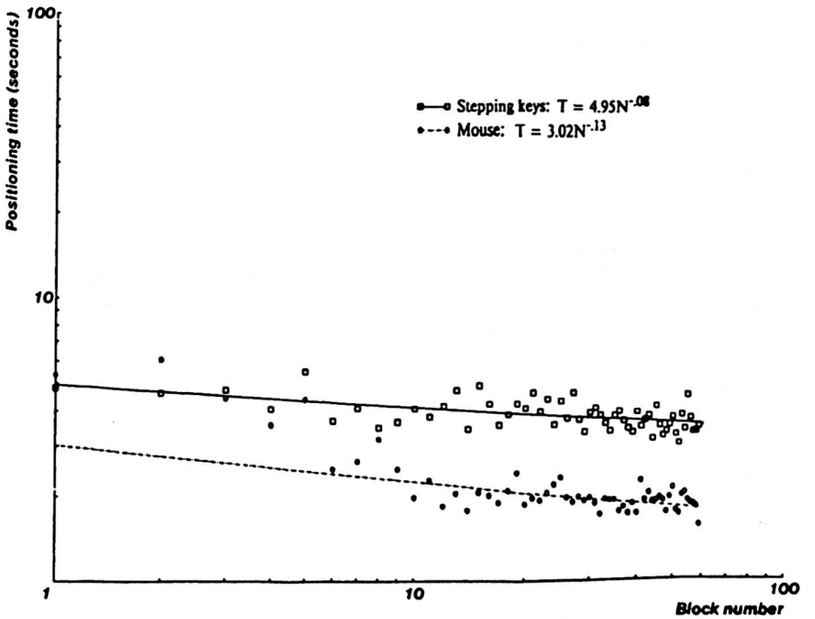

FIG. 1.5. Learning to use cursor positioning devices (log-log coordinates). Plotted from the original data for Subject 14 (Card, English, & Burr, 1978).

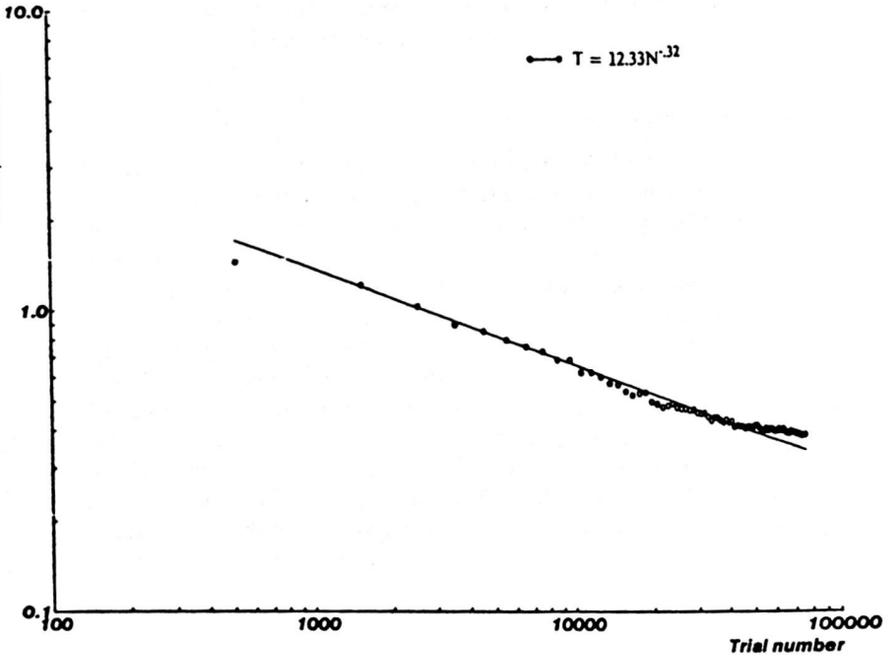

FIG. 1.6. Learning in a ten finger, 1023 choice task (log-log coordinates). Plotted from the original data for Subject JK (Seibel, 1963).

Memory

Figure 1.7 is from some unpublished work of John Anderson (1980). It shows learning performance in a task that would appear to stress mostly memory, though of course it has both a perceptual and a motor response aspect. The task is an old–new judgment on a set of simple sentences, such as "The doctor talked to the lady." There is a fixed population of grammatical subjects, objects, and verbs; a subset of these are seen initially, and then sets of the originals plus distractors (made from the same populations) are shown repeatedly. After awhile of course a subject has seen both the targets and the distractors several times. The figure shows that the reaction time to make the memory judgment follows the log-log linear law.

Complex Routines

Figure 1.8 is from some work done in connection with a general attack on understanding user–computer interaction (Moran, 1980). A specific, complex on-line editing task of completely rearranging a given sentence of three clauses is being performed repeatedly. The task is absolutely identical each time (i.e., the same sentence). Thus we are seeing a subject simply follow an internally familiar, complex plan. The top curve is the total time to perform the task. The lower

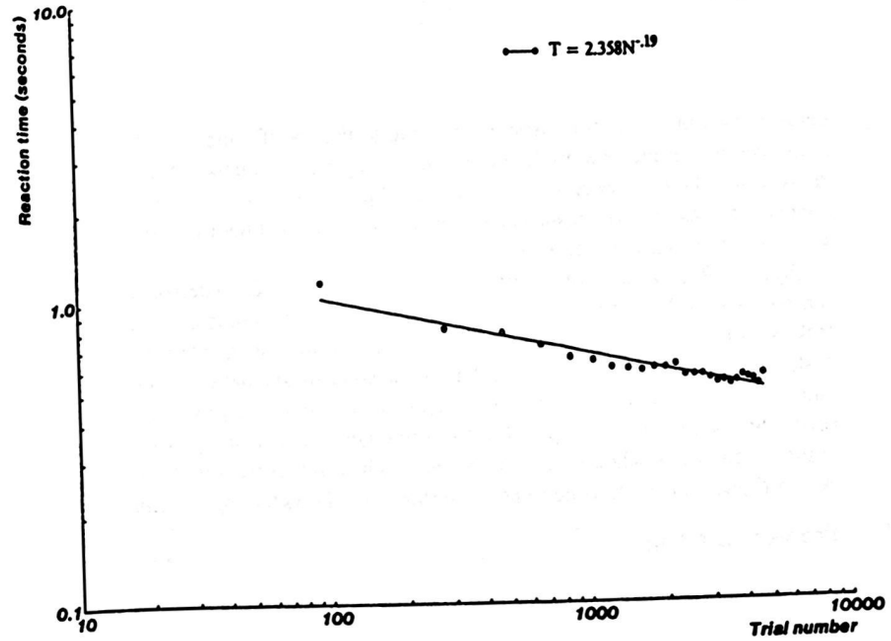

FIG. 1.7. Learning in a sentence recognition task (log–log coordinates). Plotted from the fan 1 data of Anderson (1980).

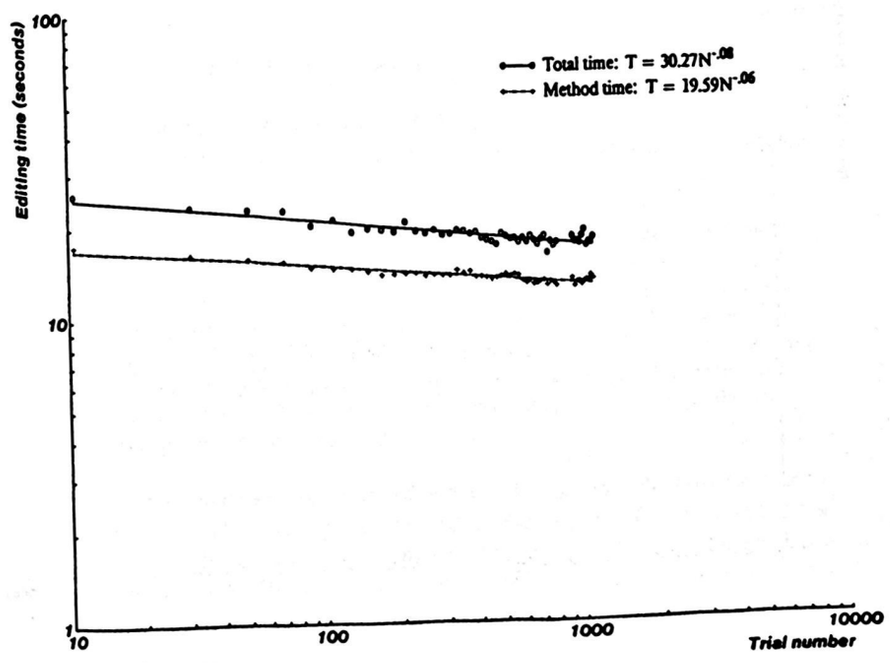

FIG. 1.8. Learning of a complex on-line editing routine (log–log coordinates). Plotted from the original data of Moran (1980).

curve shows the execution time attributable to the specific method being used, computed according to a model based on the keystroke sequence (Card, Moran, & Newell, 1980b). It decreases only if the subject makes some improvement that changes the number of keystrokes rather than decreasing think time. Both curves show log-log linear practice effects.

Figure 1.9 shows a more complex cognitive task (Neves & Anderson, in press), but one that still can be considered as evolving toward a complex routine. The task is to find the rule justifying each step in a proof in a simple formal proof system, taken to mirror the typical proof system of synthetic geometry. The subject faces a display that shows (on request) the lines of the proof, the axioms, or the theorems that are applicable to derive new steps in the proof. He must assign to each step whether it is an axiom or which rule is used in its derivation. As the figure shows, the time to perform this task follows the log-log linear law.

Problem Solving

Figure 1.10 shows our own small addition to the population of tasks known to follow the log-log linear law. As the ubiquity of the law became clear, it seemed that it was miscast as something applying only to perceptual and motor skills, but

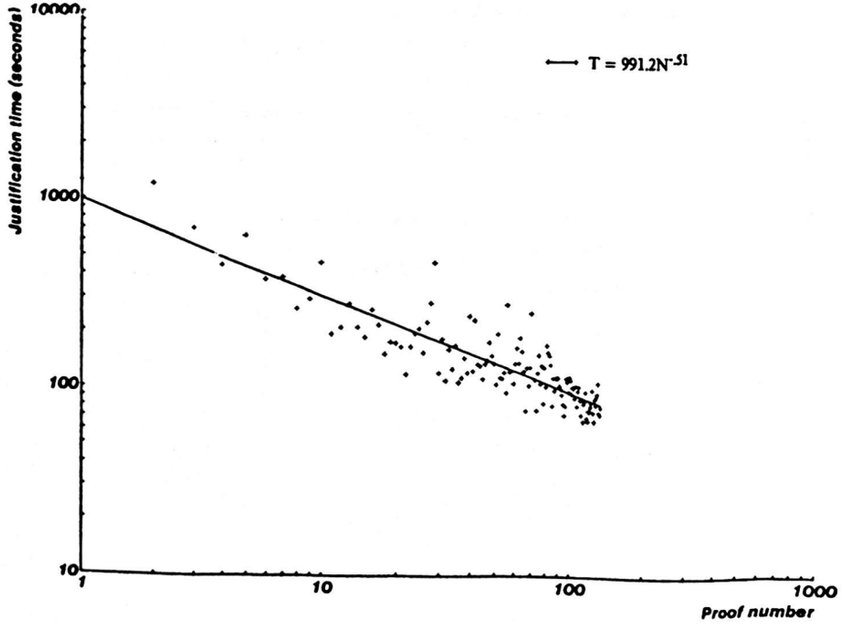

FIG. 1.9. Learning in a geometry proof justification task (log-log coordinates). Plotted from the original data (Neves & Anderson, in press).

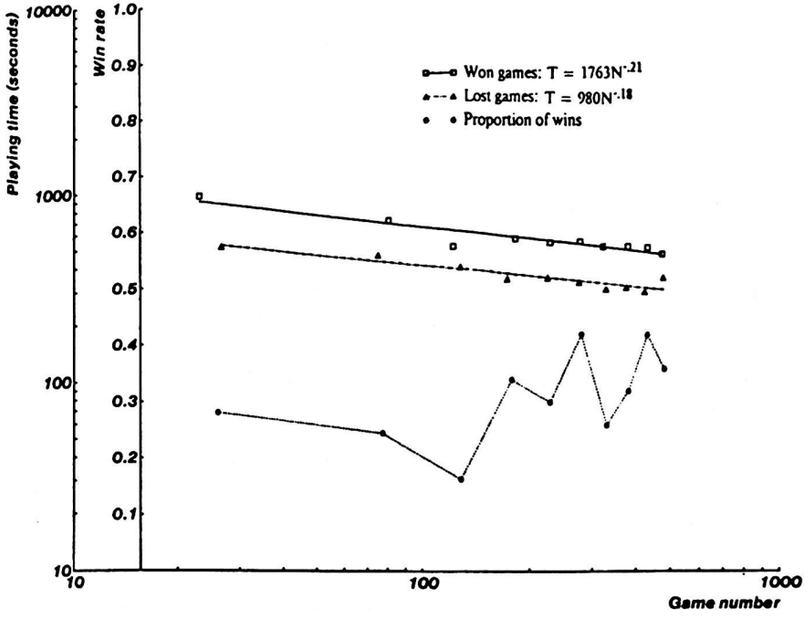

FIG. 1.10. Learning in the card game Stair (log-log coordinates).

rather it applied to all forms of mental behavior. To test whether the law applied to problem solving tasks, we had a single subject play 500 hands of a game of solitaire called *Stair*.

> Stair involves laying out all 52 cards face up from a shuffled deck, in 8 *columns* (four with 7 rows, four with 6 rows). There are also four *spots* (initially empty), each of which can hold only a single card. The aim is to build four *stacks*, ace to king, one for each suit, by moving cards around under typical solitaire constraints. A card in a spot or at the bottom of a column may be moved: (1) to a spot, if it is empty; (2) to a stack, if the card is the next in order building up; or (3) to the bottom of another column, if the card is the next lower in the same suit (e.g., the six of spades appended to the seven of spades).

The game can be seen to be one of perfect information—all cards are faceup. The shuffled deck simply picks out one of the possible initial conditions at random. From that point no further chance element enters. Whether the game can be won or not, or how many cards can be moved to the stacks, is derivable from the initial configuration. The subject, whose ability to calculate ahead is of course limited, may create a partial plan and then proceed to execute it; in doing so, he may make irrevocable moves that lose him the possibility of winning. But

such failures all arise, as in chess or checkers, because of his limited problem-solving ability. Although this task certainly has a strong perceptual component (and a weak motor component), it is to be classed as fundamentally an intellectual task, in the same way as games such as chess and checkers or problems such as the *traveling salesman problem*.

Turning to the figure, the top curve shows the time for games that the subject won; the lower curve shows the time for games that the subject lost; at the bottom the proportion of games won is shown. The points are averaged over 50 games. There is of course only one series of trials, since all games, won or lost, contribute to practice. Each group of 50 games is therefore split between the two curves before being averaged. Both curves essentially follow the log-log linear law. In general it takes longer to win than to lose, since losing involves becoming stuck after a relatively small number of cards has been played to the stack, whereas winning always involves working through all 52 cards (though the tail end goes rapidly).

The issue of the speed-accuracy tradeoff reveals itself in this data. Clearly, the subject is applying various criteria of certainty to his play. He could conceivably, as a strategy choice, study each initial layout for 5 hours before making his first move or play impulsively with no contemplation at all. In fact, the subject felt he had little genuine control of the speed-accuracy tradeoff, partly because the complexity of the initial position made it unclear whether an apparently lost game was just a bad layout or was due to a failure to spend enough time analyzing. Note that the most deviant point from the log-log line (at 150–200 trials) corresponds to the lowest win frequency.

Other Tasks and Measures

The story does not quite end at this point. Learning in other tasks and measured on other criteria seems to follow the log-log law. We give here a couple of examples.

Figure 1.11 is reproduced from Stevens and Savin (1962). It plots eight tasks with various response measures in log-log space. The criteria are all oriented to increase with practice. The plot is actually of the *cumulated* responses (i.e., the integral of the usual curve). This is just the same as the usual power law, because the integral of a power law is a power law (though integration tends to smooth the curve, helping to account for the lovely appearance of the curves, in addition to the relatively large numbers of subjects).

$$\int_1^N Bx^{-\alpha}\,dx = B(1-\alpha)^{-1}(N^{1-\alpha}-1) \qquad (3)$$

Some of these curves are time curves (actually, amount accomplished per unit time, to make them positive curves); but several are not (e.g., 1 is the number of correct anticipations in learning nonsense syllables, 2 is the time on target in a pursuit tracking task; 3 is the number of balls thrown into a target area; 4 is the number of correct responses in an animal experiment in learning a maze, and so on.)

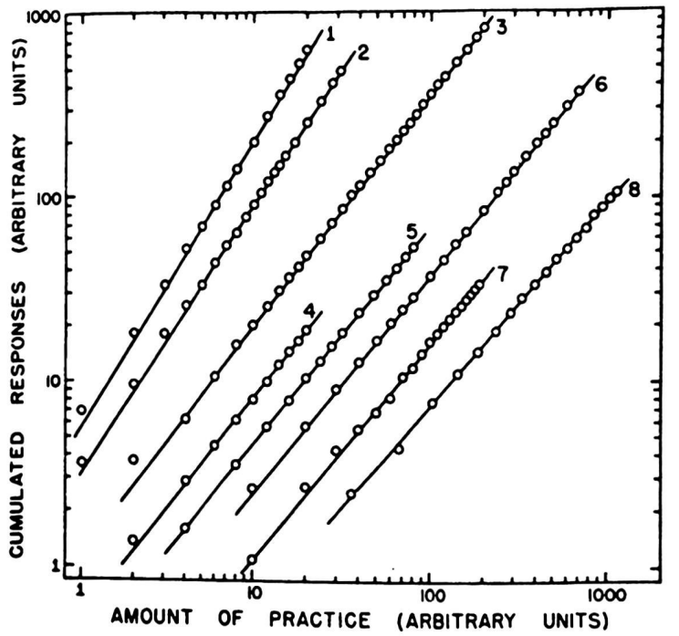

FIG. 1.11. Eight cumulated response practice curves (log-log coordinates). Figure from Stevens & Savin (1962). Copyright 1962 by the Society for the Experimental Analysis of Behavior, Inc.

As a second type of example, it has long been known in industrial engineering that the so-called learning curve for production of manufactured projects was log-log linear. In part this comes of various simple rules of thumb (e.g., "... each time the quantity of [air]planes is doubled, the cumulative average man-hours per plane will be [reduced by] 80%" [Rigon, 1944]). However, Fig. 1.12 shows an empirical curve from machine tool manufacture (Hirsch, 1952). Notice that the index of performance is not time but cost.

Summary

We have shown some 12 diverse examples of the log-log linear law of practice for trials versus time. From Table 1.1 we can make one more particular point:

- The learning rates, α, are all less than 1.

Our main point is that the law is ubiquitous when one measures the log of performance time against the log of trial number. Where the general impression

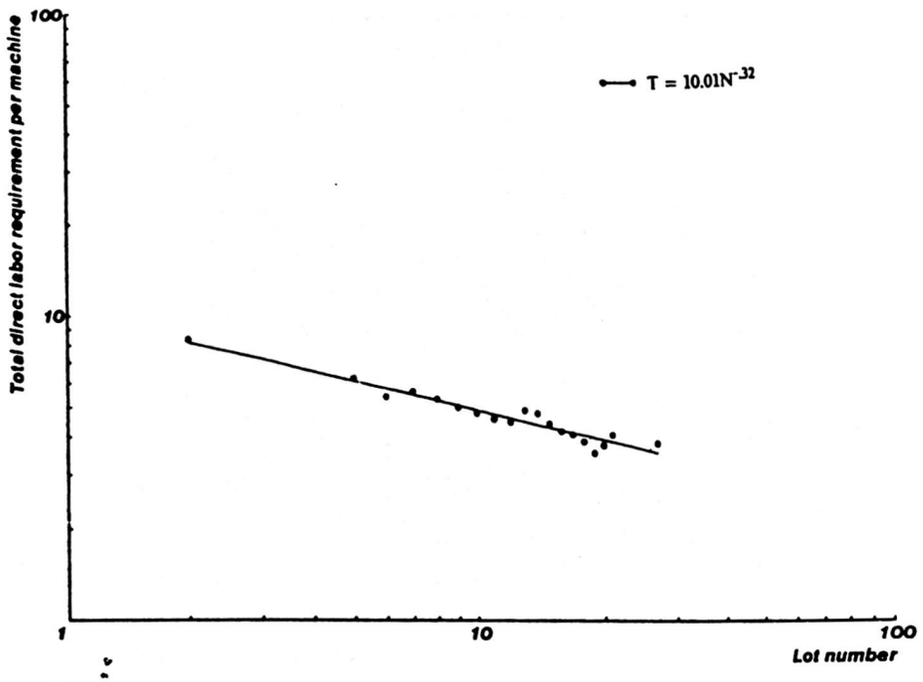

FIG. 1.12. The effect of practice on direct labor requirement in machine production (log-log coordinates). Replotted from Hirsch (1952).

seems to have been that the law showed up in perceptual-motor behavior, we think it is clear that it shows up everywhere in psychological behavior—at least it cannot easily be restricted to some part of the human operation.

Our proposition on ubiquity is extended, perhaps beyond our druthers, to learning curves involving other measures of performance and even to tasks possibly (but not certainly) beyond the pale of individual human behavior. We do not however claim that all learning is log-log linear. Nor do we claim that practice always leads to learning.

We do not wish to assert that such an effect stems from a single cause or mechanism. Indeed, its ubiquity might seem to indicate multiple explanations. We do wish to make one general comment about the regularity and what might be expected from understanding it. Its widespread occurrence implies that it depends on quite general features of the learning situation or of the system that learns. If we develop a theory that depends on detailed perceptual or motor mechanisms, we shall just create trouble for the more cognitive instances or vice versa.

One is immediately reminded of other examples of ubiquitous regularities and their explanation. The *normal distribution*, which arises out of the independent additive combination of many small increments, is the most well known.

Another, usually known as *Zipf's law*, gives the distribution for items according to their rank order, which is common to word frequencies, city sizes, incomes, and many other ordered phenomena (Simon, 1955). Consistently, highly general stochastic models underly these various phenomena. They explain the regularity but leave open the detailed mechanisms that produce the stochastic processes.

Thus, in searching for an explanation for this regularity, we should expect at best to find some such general considerations. Though it will not tell us in detail about the learning mechanism, it may still tell us something worth having.

BASICS ABOUT POWER LAWS

In this section we present some general perspectives on power laws and what they mean.

Differential Forms and Rates of Change

We start with the power law and its equivalent log-log form:

$$T = BN^{-\alpha} \tag{4}$$

$$\log(T) = \log(B) - \alpha\log(N) \tag{5}$$

It is instructive to see this in terms of the local rate of learning, dT/dN.[5]

$$\frac{dT}{dN} = -\alpha BN^{-\alpha-1} \tag{6}$$

$$= -\frac{\alpha T}{N} = -\left(\frac{\alpha}{N}\right)T \tag{7}$$

$$= -\alpha B^{-1/\alpha}T^{1+1/\alpha} \tag{8}$$

Now, one baseline form for learning is exponential. It can arise, for instance, from any mechanism that is completely local. If there is something that learns on each local part of a performance, independent of any other part, then the change in T (the sum of the changes to each part of T) is proportional to T:

$$\frac{dT}{dN} = -\alpha T \tag{9}$$

$$T = Be^{-\alpha N} \tag{10}$$

Comparing this differential form to that of the power law, shows that power-law learning is like exponential learning in which the instantaneous rate α' decreases with N, that is,

$$dT/dN = -\alpha' T \tag{11}$$

where $\alpha' = \alpha/N$

[5]For ease of exposition we treat the trial number N as a continuous variable. In fact, nothing material depends on it; we could work with finite differences throughout, at the cost of added complexity.

Both the exponential and the power function are monotonically decreasing functions that asymptote at 0. The decreasing rate of learning in the power function leads to its approaching asymptote much more slowly. Figure 1.13 shows these two curves in linear coordinates, with identical initial values ($B = 1$). This corresponds to $N = 0$ for the exponential, and $N = 1$ for the power. Thus, one way to think of power law learning is that it is a learning process in which some mechanism is slowing down the rate of learning.

Not every scheme of slowed-down learning leads to the power law. For instance, if we generalize the differential equation above we obtain a different law:

$$\frac{dT}{dN} = \left(\frac{\alpha}{N^{\beta}}\right) T, \tag{12}$$

where $\beta \neq 1$.

$$T = Be^{-\alpha N^{1-\beta}} \tag{13}$$

A representative curve for β less than 1 is also shown in Fig. 1.13, which produces asymptoting between the exponential and the power law.

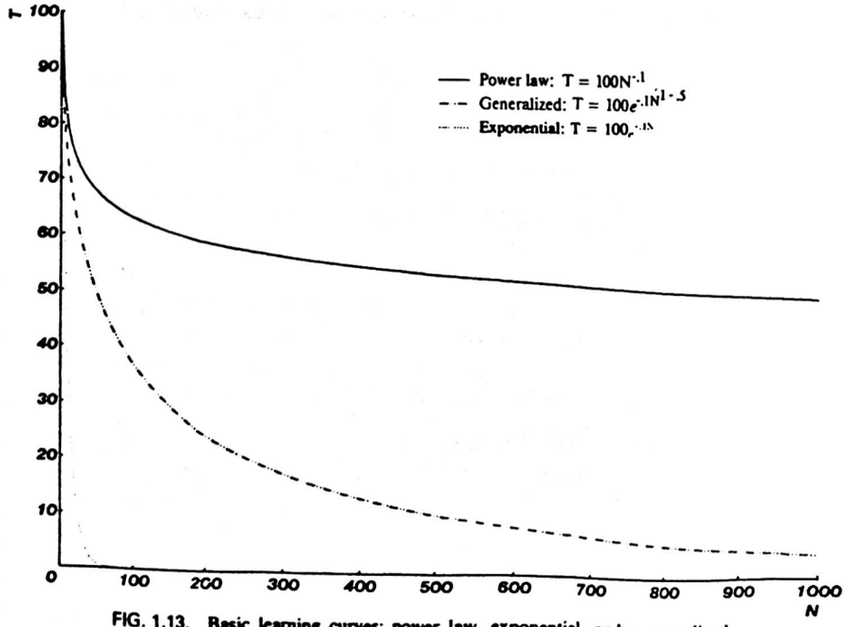

FIG. 1.13. Basic learning curves: power law, exponential, and a generalized curve.

The form of the power law can be appreciated in terms of a simple global rule, as well as in differential form:

Power Law Decay: If T decreases by a factor δ in the first N trials, it will take another $N(N - 1)$ trials to decrease by a factor of δ again.

Comparison with the corresponding global rule for the exponential, shows again how much more slowly the power law drops off:

Exponential Law Decay: If T decreases by a factor of δ in the first N trials, it will take another N trials to decrease by a factor of δ again.

Asymptotes and Prior Experience

As given in Equation 4, the law assumes: (1) the asymptote of the learning is 0 (i.e., the task can be performed in arbitrarily small time after enough learning); and (2) the initial trial of the learning occurs at the first trial of the measured series. Neither of these assumptions need be true.

The more general form of the law is

$$T = A + B(N + E)^{-\alpha} \tag{14}$$

A (≥ 0) is the *asymptote* of learning as N increases indefinitely. E (≥ 0) is the number of trials of learning that occurred prior to the first trial as measured (i.e., prior *experience*); it thus identifies the true *starting point* of learning. (Neither $A < 0$ or $E < 0$ make immediate sense, given these interpretations; $A = 0$, $E = 0$ reproduces the basic form of Equation 4.)

Plotting $\log(T - A)$ against $\log(N + E)$ still yields a straight line whose slope is $-\alpha$. The difficulty of course is that A and E are not known in advance, so the curve cannot be plotted as an initial exploratory step in an investigation.

One alternative is just to plot in $\log(T) - \log(N)$ space and understand the deviations:

$$\log(T - A) = \log(B) - \alpha\log(N + E) \tag{15}$$

$$\log(T) = \log(B) - \log(1 - A/T) - \alpha\log(N) - \alpha\log(1 + E/N) \tag{16}$$

There is an error term for each parameter. If T is large with respect to the asymptote, A, then $\log(1 - A/T)$ is close to $\log(1)$, which is 0. This occurs at early values of N. If N is large with respect to E, then $\log(1 + E/N)$ is close to $\log(1)$, which is 0. Thus, the two deviations affect the curve at opposite parts: Non-zero values of E distort the straight line for low N, non-zero values of A distort it for high N.

Figure 1.14 shows a power law with a starting point ($-E$) of -25 and a time asymptote (A) of 5. Figure 1.15 shows the same curve in log-log space. Characteristically, the starting point pulls the initial segment of the curve down toward the horizontal and the finite asymptote pulls the high N tail of the curve up toward the horizontal. A central region of the curve appears as a straight line. It is however less than the true slope ($-\alpha$), as the line shows.

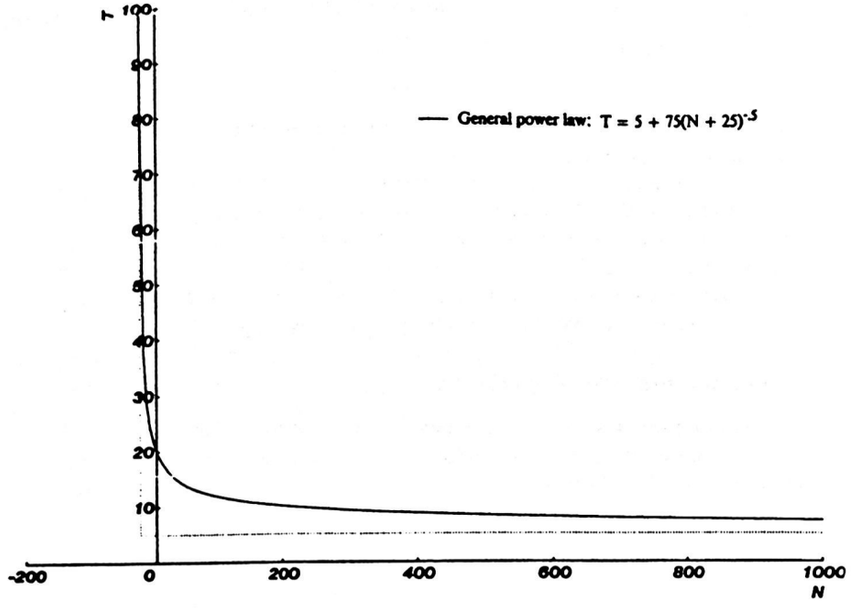

FIG. 1.14. A general power law curve.

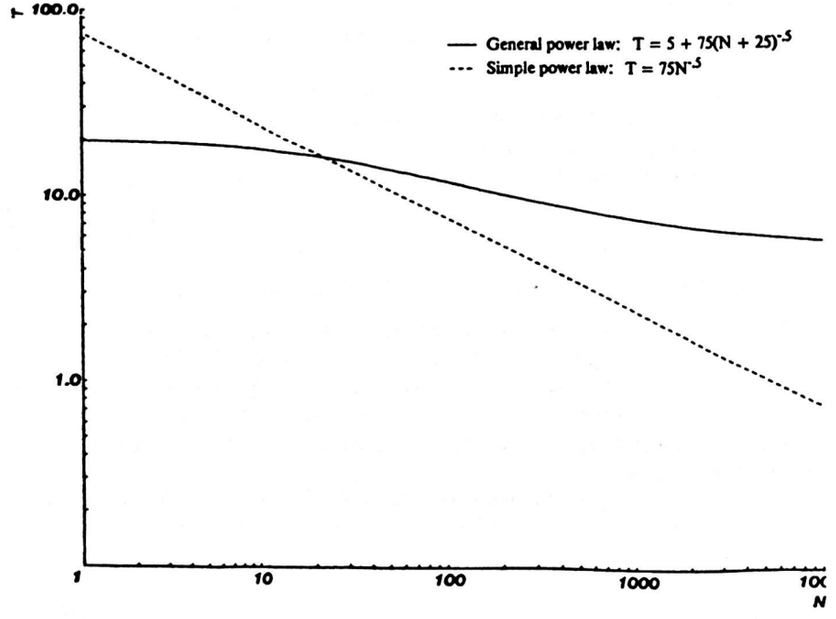

FIG. 1.15. A general power law in log-log coordinates. The simple power law with the same α and B is also shown.

The derivative of the general power function in log-log space is given by

$$\frac{d[\log(T)]}{d[\log(N)]} = -\alpha\left(1 - \frac{A}{T}\right)/\left(1 + \frac{E}{N}\right) \tag{17}$$

It can be seen that the slope is everywhere smaller than α and becomes increasingly so as either A or E increases. A reasonable estimate of the apparent slope as viewed on the graph, α^*, is at the inflection point. It is easy to obtain by setting the derivative of Equation 17 to zero:

$$\frac{d}{dN}\left[\frac{d[\log(T)]}{d[\log(N)]}\right] = -\left(\frac{\alpha}{N}\right)\left(\frac{E}{N} - \frac{\alpha A}{T}\right)\left(1 - \frac{A}{T}\right)\left(1 + \frac{E}{N}\right)^{-2} = 0 \tag{18}$$

$$\alpha^* = \frac{(\alpha N^* - E)}{(N^* + E)} \tag{19}$$

N^* is the point at which the inflection occurs. The exact value of N^* is not expressible in simple terms, but a reasonable approximation is

$$N^* = \left[\frac{BE}{\alpha A}\right]^{1/(1+\alpha)} \tag{20}$$

where $E/N^* << \alpha < 1$.

The structure of Fig. 1.15 suggests that many of the deviations in the empirical curves could be due simply to starting point or asymptote effects. Because the effect of these two phenomena is to bend toward the horizontal at separate ends, it is possible to tell from the curve in log-log space what effect might be operating. The original Snoddy data in Fig. 1.1 provides an example of a clear initial deviation. It cannot possibly be due to an earlier starting point, because the initial curve rises toward the vertical. However, it could be due to the asymptote, because raising the asymptote parameter (A) will pull the right-hand part of the curve down and make its slope steeper. The Seibel data in Fig. 1.6 provides an example where there are deviations from linearity at both ends. Use of a nonzero value for E (previous experience) will steepen the initial portion of the curve, whereas doing likewise for A will steepen the high N portion of the curve. (The results of such a manipulation are seen in Fig. 1.21.)

Trials or Time?

The form of the law of practice is performance time (T) as a function of trials (N). But trials is simply a way of marking the temporal continuum (t) into intervals, each one performance-time long. Since the performance time is itself a monotone decreasing function of trial number, trials (N) becomes a nonlinear compression of time (t). It is important to understand the effect on the law of practice of viewing it in terms of time or in terms of trials.

The fundamental relationship between time and trials is

$$t(N) = T_0 + \sum_{i=1}^{N} T_i = T_0 + \sum_{i=1}^{N} Bi^{-\alpha} = T_0 + B \sum_{i=1}^{N} i^{-\alpha} \qquad (21)$$

T_0 is the time from the arbitrary time origin to the start of the first trial. This equation cannot be inverted explicitly to obtain an expression for $N(t)$ that would permit the basic law (Equation 4) to be transformed to yield $T(t)$. Instead, we proceed indirectly by means of the differential forms. From Equation 21 we obtain

$$\frac{dt}{dN} = T \qquad (22)$$

Think of the corresponding integral formulation,

$$\frac{d}{dz} \int_a^z f(x)\, dx = f(z)$$

Now, starting with the power law in terms of trials we find:

$$\frac{dT}{dt} = \frac{dT/dN}{dt/dN} = \frac{-\alpha T/N}{T} = \frac{-\alpha}{N} \qquad (23)$$

But from the basic Equation (4):

$$N = \left(\frac{T}{B}\right)^{-1/\alpha} \qquad (24)$$

Thus, we obtain the trials power law reexpressed in terms of time:

$$\frac{dT}{dt} = -\alpha B^{-1/\alpha} T^{1/\alpha} \qquad (25)$$

For $\alpha \neq 1$ this integrates to yield

$$T^{-(1-\alpha)/\alpha} = (1 - \alpha)B^{-1/\alpha} t + C \qquad \text{for } \alpha \neq 1 \qquad (26)$$

But C is an arbitrary constant of integration and if the origin and scale of t is adjusted appropriately, we find:

$$T = B' t^{-\alpha/(1-\alpha)} \qquad \text{for } \alpha \neq 1 \qquad (27)$$

Thus, a power law in terms of trials is a power law in terms of time, though with a different exponent, reflecting the expansion of time over trials. The results are significantly altered when $\alpha = 1$ (the hyperbolic) however. Equation 25 becomes

$$\frac{dT}{dt} = -B^{-1} T \qquad (28)$$

This is no longer the differential form of a power law. Instead it is that of an exponential:

$$T = Ce^{-B^{-1}t} \qquad (29)$$

It is left as an exercise for the reader to confirm that an exponential function in trials transforms to a *linear* function in time (hence, Zeno-like, an infinite set of trials can be accomplished in a finite amount of time).

FITTING THE DATA TO A FAMILY OF CURVES

Given empirical curves, such as occur in abundance in the second section, it is important to understand how well they are described by curves of a given family (e.g., power laws) and whether there are alternative general forms that fit them just as well (as noted in the introduction, exponential, hyperbolic, and logistic curves have enjoyed much more favor than power functions). Curve fitting without benefit of a model is notoriously a black art. Nonetheless, we have deliberately chosen not to be model driven initially, because we want to have empirical generalizations as the starting point in the search for theory, not just the raw data.

The basic issue of curve fitting can be introduced from Seibel's own treatment of his data (Fig. 1.6), which appears to be an extremely good fit to the log-log law over an extensive range (40,000 trials). Seibel (1963) fit his points to three curves by least squares: (1) a power law with asymptote only (i.e., E fixed at 0); (2) an exponential with asymptote; and (3) a general power law with both asymptote and starting point.[6] He obtained an r^2 of .991 for the power function with asymptote only. But he also obtained an r^2 of .971 for the exponential with asymptote. His general power law fit was .997. (His parameters for asymptotes and starting points are mostly reasonable but not entirely.) Thus, all the curves give good fits by normal standards. If only differences in the least-squared residual are used, there can hardly be much to choose from. This is an annoying result, in any case; but it is also somewhat unexpected, for the plots that we have shown, though they surely contain noise, are still impressively linear by intuitive standards and involve lots of data.

It is important to recognize that two basic kinds of failure occur in fitting data to a family of smooth curves: (1) failure of the shape of the data curve to fit to the shapes available within the family; and (2) noise in the data, which will not be fit by any of the families under consideration or even noticeably changed by parametric variation within a family. These distinctions are precisely analogous to the frequency spectrum of the noise in the data. However, the analogy probably should not be exploited too literally, because an attempt to filter out the high-frequency noise prior to data fitting simply adds another family of empirical curves (the filters) to confound the issues. What does seem sensible is to attempt to distinguish fits of shape without worrying too much about the jitter.

A simple example of this point of view is the (sensible) rejection of the family of logistic curves from consideration for our data. The logistic provides a sig-

[6]The exponential is translation invariant, so a special starting point is not distinguishable for it; that is, $Be^{N+E} = (Be^{E})e^{N} = B'e^{N}$.

moid curve (i.e., a slow but accelerating start with a point of inflection and then asymptoting). No trace of an S-shape appears in any of our data, though it would not be lost to view by any of the various monotone transformations (logs, powers, and exponentials) that we are considering. Hence, independent of how competing the measure of error, the logistic is not to be considered.

The size of the jitter (i.e., the high-frequency noise) will limit the precision of the shape that can be detected and the confidence of the statements that can be made about it. It provides a band through which smooth curves can be threaded, and if that band is wide enough—and it may not have to be very wide—then it may be possible to get suitable members of conceptually distinct curves through it. In all cases, the main contribution to any error measure will be provided by the jitter, so that only relatively small differences will distinguish the different families.

The Data Analysis Procedure

With the elimination of the logistic from consideration, we have focused our efforts on three families of curves: *exponential*, *hyperbolic*, and *power law*. The analysis procedure that we have ended up using is primarily graphical in nature. We look at what types of deviations remain, once an empirical curve has been fit optimally by a family of theoretical curves. The analysis consists of judgments as to whether the deviations represent actual distortions of shape, or merely jitter. The procedure has the following components:

1. Find spaces where the family of curves should plot as straight lines. Judgments of shape deviation are most easily made and described when the norm is a line. These are the *transformation spaces* of the given family. There may be more than one such space.
2. For each family of curves, find the best linear approximation to the data in the transformation spaces of the family. This will generally involve a combination of search and linear regression.
3. Accept a curve for a family, if the best fit plots as a straight line in the space of that family. Reject it, if it has significant shape distortion.
4. Understand the shape distortion of family X when plotted in the space of family Y. Expect curves of family X to show the characteristic distortion when plotted in the spaces of alternative families.
5. Compute an estimate of fit (r^2) for the best approximation in each transformation space. Expect these values to support the judgments made on the basis of shape distortion.

These criteria contain elements both of acceptance and rejection and provide a mixture of absolute judgments about whether data belong to a given family and relative judgments about the discrimination between families. The parameters for the best fits as well as the estimates of fit (r^2) can be found in Table 1.2.

TABLE 1.2
The General Learning Curves: Parameters from Optimal Fits in the *Log* Transformation Spaces

	Exponential $T = A + Be^{-\alpha N}$				*Hyperbolic* $T = A + B/(N + E)$				*Power Law* $T = A + B(N + E)^{-\alpha}$				
Data Set	*A*	*B*	α	r^2	*A*	*B*	*E*	r^2	*A*	*B*	*E*	α	r^2
Snoddy (1926)	27.01	38.80	.061	.916	24.49	243.6	1.3	.962	21.74	119.2	0.0	.71	.975
Crossman (1959)	7.19	4.59	1×10^{-7}	.842	7.10	2.4×10^{6}	151000	.983	6.91	20481	31000	.66	.990
Kolers (1975) - Subject HA	1.36	3.82	.018	.849	1.10	94.02	9.8	.915	.18	15.25	0.0	.46	.931
Neisser et al. (1963)													
Ten targets	.06	.83	.13	.905	.00	2.74	.9	.965	.00	2.35	.6	.95	.965
One target	.06	.44	.094	.938	.00	3.16	4.6	.951	.00	2.57	3.9	.94	.951
Card, English & Burr (1978)													
Stepping keys - Subj. 14	2.35	1.99	.011	.335	2.14	171.4	75.2	.338	.02	6.36	9.3	.14	.340
Mouse - Subj. 14	1.46	1.28	.028	.452	1.46	16.70	5.0	.603	.59	4.28	0.0	.33	.729
Seibel (1963) - Subject JK	.371	.461	.000055	.956	.328	3888.1	3042	.993	.324	2439.9	2690	.95	.993
Anderson (1980) - Fan 1	.487	.283	.00055	.774	.466	231.6	319.7	.902	.353	4.322	0.0	.39	.947
Moran (1980)													
Total time	13.80	6.66	.00073	.546	14.77	3335.9	474.6	.637	.03	30.24	0.0	.08	.839
Method time	11.61	3.11	.0010	.652	11.75	1381.8	360.0	.737	.26	19.35	0.0	.06	.882
Neves & Anderson (in press)													
Total time - Subject D	57.5	240.2	.019	.660	45.6	5000.2	7.3	.728	0.0	991.2	0.0	.51	.780
The Game of Stair													
Won games	476	319	.0052	.689	449	29800	40.1	.783	120	1763	0.0	.25	.849
Lost games	152	326	.0016	.634	247	41270	124.1	.751	1	1009	2.5	.19	.841
Hirsch (1952)	2.76	4.35	.070	.819	2.34	37.05	4.9	.897	.00	10.01	0.0	.32	.932
General Power Law													
$T = 5 + 75(N + 25)^{-0.5}$	7.21	6.78	.0037	.983	6.41	1069.6	91.2	.997	5.00	74.85	24.9	.50	1.000
40 Term Additive Mixture	1.60	45.37	.0065	.904	.58	1231.2	10.2	.997	.19	753.1	7.2	.89	.998
Chunking Model													
Combinatorial TE	4.61	4.71	.0046	.957	4.35	365.7	55.3	.992	2.86	17.40	6.6	.33	1.000

The remainder of this section shows how we applied this data analysis procedure. We start by looking at the transformation spaces. This is followed by an examination of the distortions that occur when a theoretical curve is plotted in a space belonging to a different family. We are then in a position to analyze a couple of the empirical curves that appeared in the second section.

The Transformation Spaces

The curves that we are interested in belong to multiparameter families (3 for the exponential and hyperbolic; 4 for the power law). Regression can be used to fit a line to an empirical curve plotted in a multidimensional space. Unfortunately, for the three families that we are interested in, there is no space in which all the parameters (three or four) can be determined by linear regression. The most that we can obtain is two parameters. The remainder must be determined by some other means, such as search. The choice of which parameters are to drop out of the analysis determines the transformation space. We have primarily worked in two different types of transformation spaces. The first type consists of the *log* spaces. These are the most commonly used linearizing spaces for functions with powers. The log transformations that we use are the following:

$$\text{Exponential: } T' = \log(B) - \alpha N \quad \text{for } T' = \log(T - A) \tag{30}$$

$$\text{Hyperbolic: } T' = \log(B) - N' \quad \text{for } T' = \log(T - A) \text{ and } N' = \log(N + E) \tag{31}$$

$$\text{Power Law: } T' = \log(B) - \alpha N' \quad \text{for } T' = \log(T - A) \text{ and } N' = \log(N + E) \tag{32}$$

The log spaces for the hyperbolic and the power law turn out to be the standard log-log space, whereas the exponential is in semilog space. Determining fits in these spaces requires a combination of search (over $0 \leq A \leq T_{\min}$ and $0 \leq E$) and regression (for B and α). Because the exponential and hyperbolic families are each missing one of these parameters, the process becomes simpler for them. The exponential only requires a one-dimensional search (over $0 \leq A \leq T_{\min}$), whereas the hyperbolic can replace the regression (for B and α) with the computation of the average for B.

The log spaces have been used exclusively for the data analyses that are described in the following section (Table 1.2 was computed in the log spaces). It is important to realize though that they are not the only transformation spaces that can be used. We have explored what we call the T-X spaces, though space precludes presenting the analysis. Transforming a curve into its T-X space involves pushing all the nonlinearities into the definition of X as follows:

$$\text{Exponential: } T = A + BX \quad \text{for } X = e^{-\alpha N} \tag{33}$$

$$\text{Hyperbolic: } T = A + BX \quad \text{for } X = \frac{1}{(N + E)} \tag{34}$$

$$\text{Power Law: } T + A + BX \quad \text{for } X = (N + E)^{-\alpha} \tag{35}$$

In the T-X spaces, searches are over $\alpha \geq 0$ and $E \geq 0$, with A and B determined by regression. Only single-dimensional searches are needed for the two three-parameter families. The T-X spaces prove especially useful for estimating the asymptote (A), because it maps into the intercept of the transformed curve.

The Theoretical Curves

When a curve is optimally fit in a space corresponding to its family, it plots as a straight line (by definition). This is not true though when the curve is fit in a space corresponding to some other family. There will be distortions that show up as nonlinearities in the plot. By understanding these characteristic shape distortions, we are able to interpret the deviations that we find when we plot the data in these spaces. This will help us to distinguish between random jitter and distortions that signal a bad fit by the family of curves. Data that plot with the same deviations as one of the theoretical curves have a good chance of belonging to that curve's family.

Figure 1.16 shows the best that a power law can be fit in exponential log space. The power law curve is

$$T = 5 + 75(N + 25)^{-0.5} \tag{36}$$

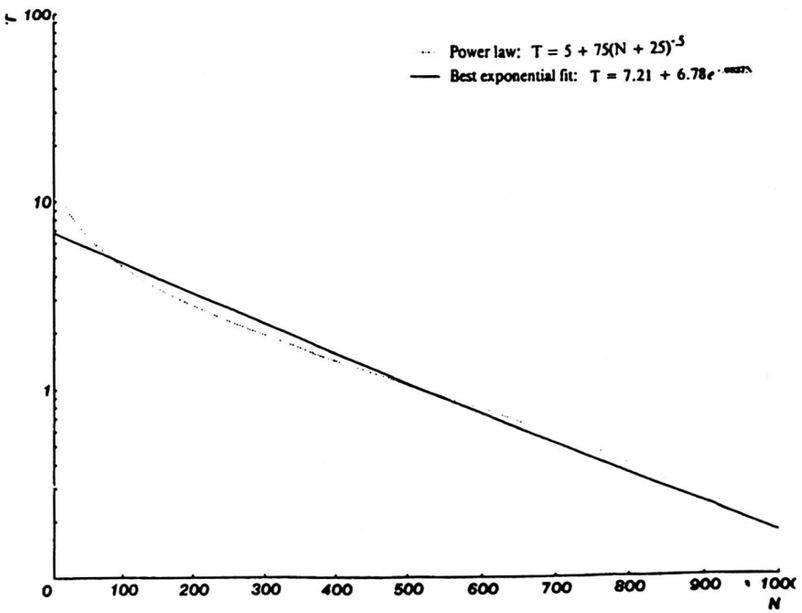

FIG. 1.16. Optimal fit of a power law in the exponential transformation space (semi-log coordinates).

This is the same curve that is plotted in Figs. 1.14 and 1.15. The parameters for the optimal exponential fit can be found in Table 1.2. The r^2 value of .983 is deceptively high, as an examination of Fig. 1.16 shows. There are strong deviations in all portions of the curve. The curve starts out high, goes low, then high again, and finally tails off downward. If we see deviations of this type when a set of data has been optimally fit by an exponential, we can conclude that the exponential family is not a good model for the data and that the power law might be.

Figure 1.17 shows the same curve optimally fit in hyperbolic log space. We see the same sorts of deviations that were found in the exponential case, but they are much attenuated. It will be hard to rule out the hyperbolic family in such a case because the variability of the data is likely to swamp out much of the distortion. At most we can hope to see the slight upturn at low N and the slight downturn for high N.

It is not necessary to look at the theoretical plots for the hyperbolic, as it is a special case of the power law. It will plot with no distortion in the power law log space, and it will have the same type of distortion in the exponential log space as did the power law. This leaves only exponential curves to be examined. We cannot present a plot of the optimal fit of an exponential in the power law log space. All attempts to find such optimal fits have led to at least one of the

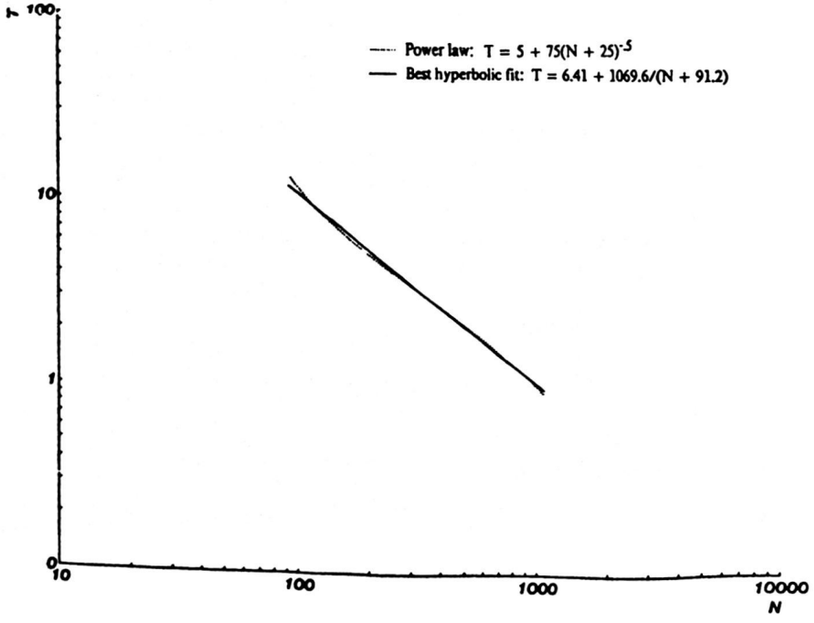

FIG. 1.17. Optimal fit of a power law in the hyperbolic transformation space (log-log coordinates).

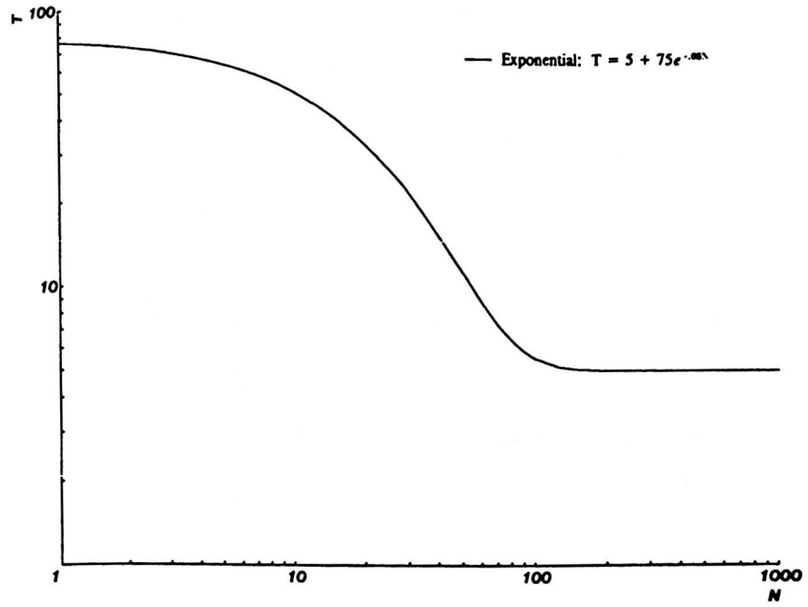

FIG. 1.18. A general exponential function in log-log coordinates.

parameters requiring a value that is too large to be represented in our computer. Though this makes the generation of a plot impossible, this information can be used in lieu of a plot. If analysis in the power law log space leads to immense parameter values, then that is evidence against a power law and for an exponential.

In addition to this information, it is useful to see what an exponential function looks like in log-log space. Figure 1.18 is characteristic of such plots. In log-log space, exponentials tend to have a flat portion followed by a rapid drop to asymptote. The central portion is considerably steeper ($\alpha > 1$) than the equivalent portion of the empirical curves that we have seen, and the asymptote is approached more suddenly.

The Analysis of a Data Set

We can now use the machinery that we have generated to analyze the data from some of the tasks in the second section. There is no space to provide a detailed examination of the data analysis techniques or of their results over the entire data set. But we do need to illustrate them enough to support the conclusions. To do this we look closely at two curves: Kolers's subject 3 (Fig. 1.3) and Seibel's subject JK (Fig. 1.6).

We first attempt to show that the exponential is not a good fit to the data, that shape distortions remain, even though the measure of fit is impressive. Then we attempt to show that both the general power and the hyperbolic families provide adequate representations of the empirical curves.

The Exponential Family

Figure 1.19 shows the optimal fit of Seibel's data in the exponential log space. As was true of the theoretical power-law curve, the value of r^2 and the plot of the optimal fit tell different stories. The value of r^2 is a respectable .956, so the exponential family can account for over 95% of the variance of Seibel's data. The characteristic power-law distortions can be clearly seen in the figure though. The value of r^2 notwithstanding, Seibel's data is not adequately fit by an exponential curve.

The same distortions can be seen in Kolers's data when it is optimally fit by an exponential (Fig. 1.20). Though they are somewhat obscured by the variability of the data, there are significant nonlinearities. With respect to the optimal fit, the data is high, then low, then high, and finally low again. These distortions are the signal that Kolers's data is also not adequately fit by an exponential curve.

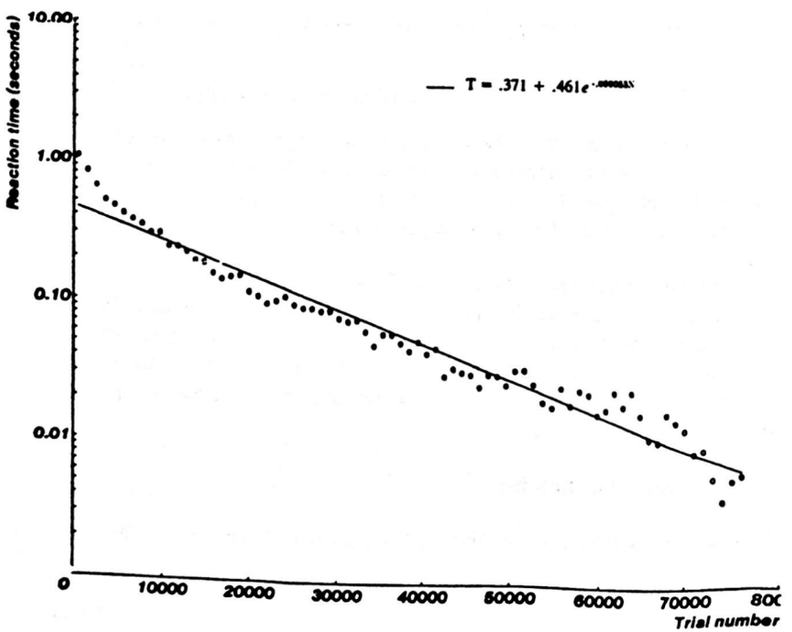

FIG. 1.19. Optimal fit to Seibel's data in the exponential transformation space (semi-log coordinates).

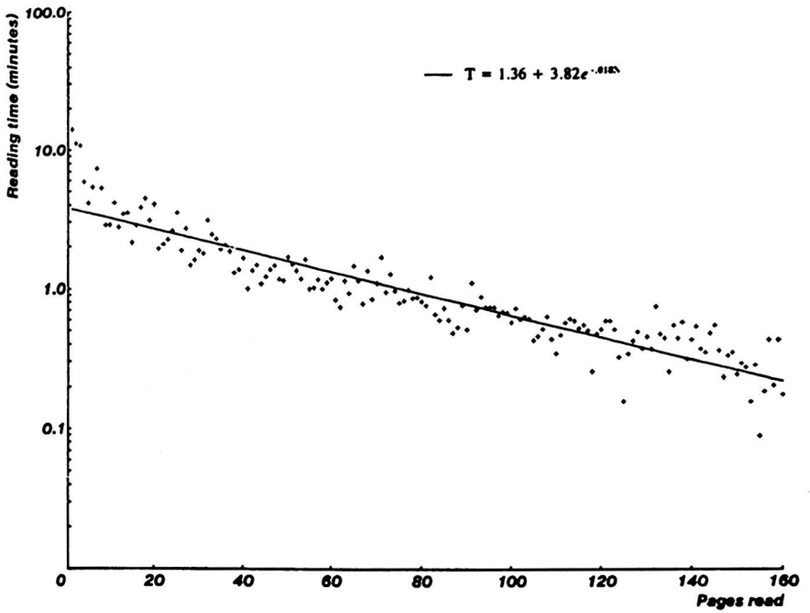

FIG. 1.20. Optimal fit to Kolers's data in the exponential transformation space (semi-log coordinates).

The Power-Law Family

In contrast to the exponential plots, the power-law plots are highly linear. Figures 1.21 and 1.22 show the optimal power-law transformations for the two data sets. Very little needed to be done to Kolers's data to achieve the optimal fit (the asymptote was assigned the value of .18). There was not much to straighten out in Kolers's data to begin with. Figure 1.3 shows that even the raw log–log plot of the data is quite linear. Seibel's data is a different matter though. In the raw log–log plot it has deviations at both ends of the curve. By giving non-zero values to the asymptote (.324) and to the prior experience (2690), the data gets straightened. This straightening yields a sharply higher α. It rises from .32 to .95 during this process. Though seemingly large, the initial experience of 2690 trials is not excessive, given the full trial range of 70,000.

The linearity of the optimal power-law plots is strong evidence for the power law as a model of learning curves. This is bolstered even further by the r^2 values that are considerably higher than those for the equivalent exponential fits (.993 versus .956 for Seibel and .931 versus .849 for Kolers). An examination of Table 1.2 reveals that the value of r^2 for a power law fit is higher than for an exponential fit for all of the practice curves that we have examined.

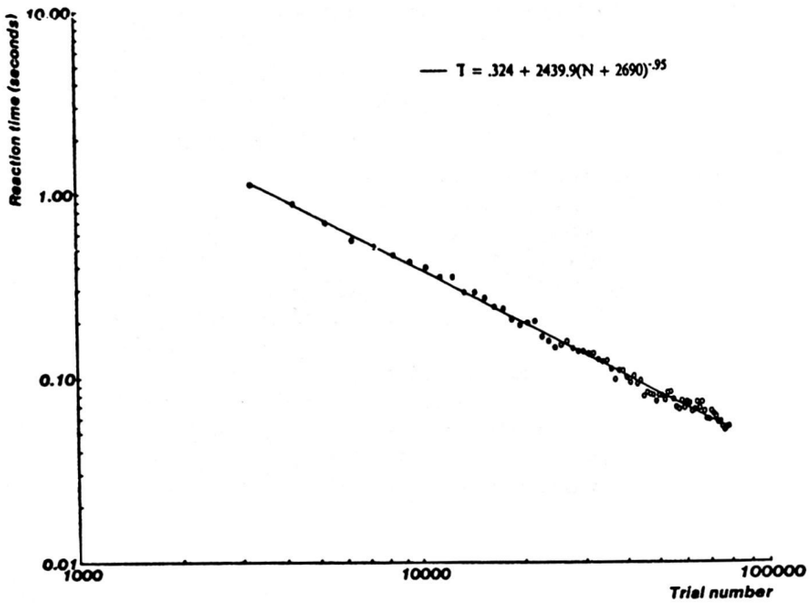

FIG. 1.21. Optimal fit to Seibel's data in the power law transformation space (log-log coordinates).

The Hyperbolic Family

It is not surprising that Seibel's data is well fit by a hyperbolic as the optimal power (α) turned out to be .95. The r^2 value remains unchanged in a shift of α to 1, and the plot remains highly linear (Fig. 1.23). What is more surprising (considering the amount of data involved) is that Kolers's data (with an optimal α of .46) is also adequately fit by a hyperbolic (Fig. 1.24). By assuming larger values for A and E, the whole curve is tilted to be steeper. There is a small loss in r^2, from .931 for the power law to .915 for the hyperbolic, but it is nowhere near as large a drop as to the exponential (.849). There does appear to be a small upturn at the beginning of the curve, and a similar downturn at the end, but the overall deviation from linearity is not large. This small inferiority of the hyperbolic (with respect to the power law) must be traded off against the fact that it has one less parameter.

Summary

Table 1.2 shows the results of this analysis for all the data sets shown in the second section. We believe that it establishes the reasonableness of excluding the possibility that practice learning is exponential and the reasonableness of describ-

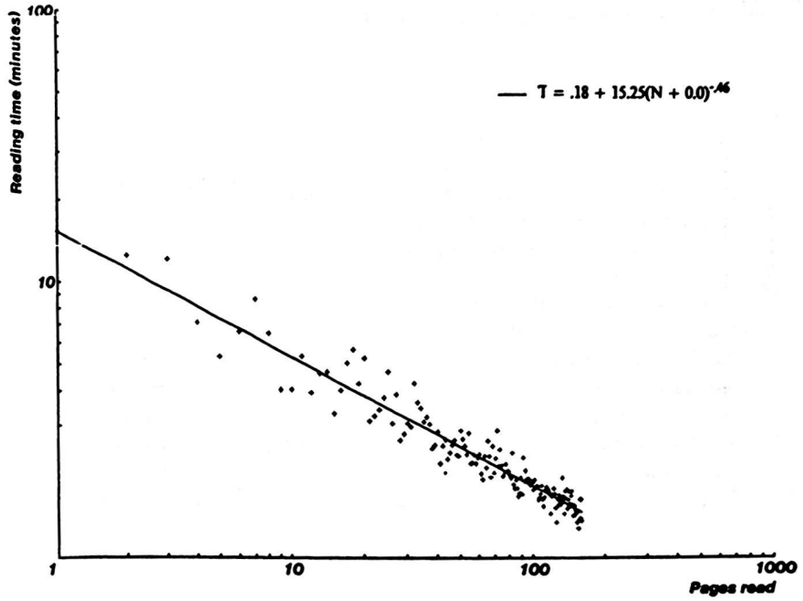

FIG. 1.22. Optimal fit to Kolers's data in the power law transformation space (log-log coordinates).

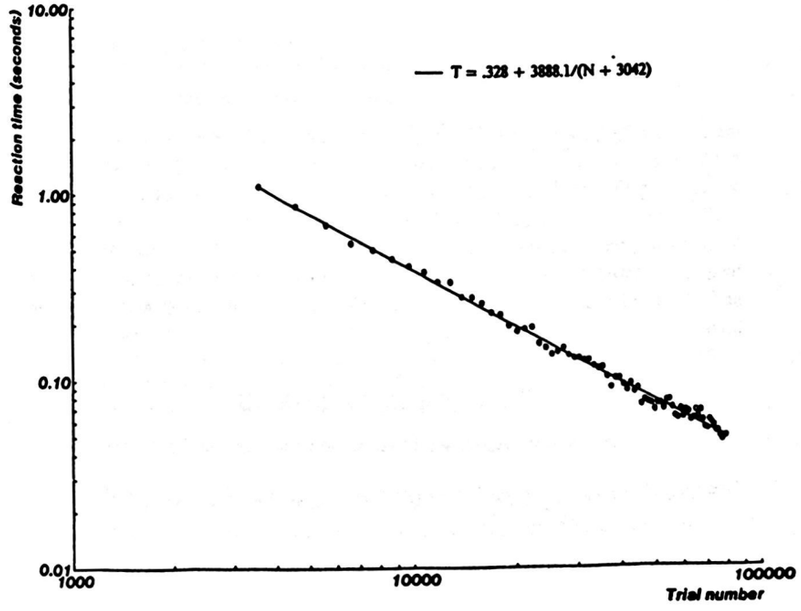

FIG. 1.23. Optimal fit to Seibel's data in the hyperbolic transformation space (log-log coordinates).

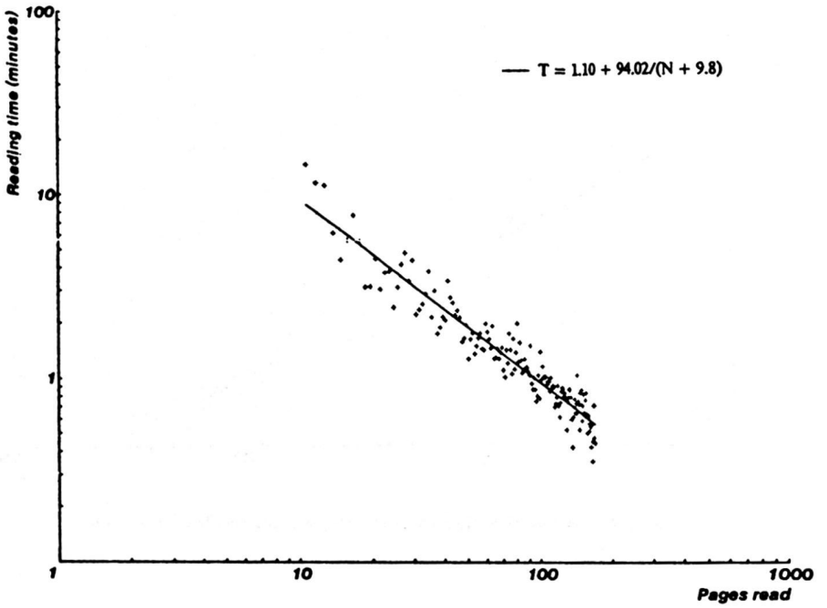

FIG. 1.24. Optimal fit to Kolers's data in the hyperbolic transformation space (log-log coordinates).

ing the data by power laws. The hyperbolic family is somewhere in the middle. From Table 1.2 it is apparent that most of the data sets can be adequately modeled as hyperbolics. There are cases though, such as the data from Moran (1980), that do seem to suffer by the loss of the extra parameter. It would be nice to be more precise about the appropriateness of the hyperbolic, but the data we have considered do not allow it. These conclusions agree with those of Mazur and Hastie (1978) in rejecting exponentials but not in rejecting general power laws.

POSSIBLE EXPLANATIONS

For the purposes of this paper, we have come to accept two propositions:

- Practice learning is described by performance-time as a power function of the number of trials since the start of learning (the hyperbolic is included as a special case).
- The same law is ubiquitous over all types of mental behavior (possibly even more widely).

What are the possible explanations for such a regularity? In this section we try to enumerate the major alternatives and to concentrate on one.

There seem to be three major divisions of explanation. The first reaches for the most general characteristics of the learning situation, in accord with the end of the second section that such a widespread phenomenon can only result from some equally widespread structural feature. One of the assumptions underlying much of cognitive psychology is the *decomposability* of thought processes. A task can be broken down into independent subtasks. *Mixture* models attempt to derive the power law from the aggregate behavior of such a collection of independent learners. The second division is some sort of improving statistical selection, in the manner of mathematical learning theory or evolution. No specific orientation exists to obtain the power law. Rather, simple or natural selective schemes are simply posited and examined. The third division takes the exponential as somehow the *natural* form of learning. Observing that the power law is much slower, it seeks for what slows down learning. What could be *exhausted* that keeps the learning from remaining exponential?

We shall concentrate on an explanation of the exhaustion type. However, we do not consider it the exclusive source of the power law of practice. So we first wish to lay out the wider context before narrowing to one.

General Mixtures

The following qualitative argument has a certain appeal.

> *The Mixtures Argument:* Performance depends on a collection of mechanisms in some monotone way [i.e., an increase in the time taken for any mechanism increases (possibly leaves unchanged) the total performance time]. The learning mechanisms that improve these performance mechanisms will have a distribution of rates of improvement—some faster, some slower. At any moment total system learning will be dominated by the fast learners, since a fortiori they are the fast ones. However, the fast learners will soon make little contribution to changes in total performance, precisely because their learning will have been effective (and rapidly so, to boot), so the components they affect cannot continue to contribute substantially to total performance. This will leave only slow learners to yield improvement. Hence the rate of improvement later will be slower than the rate of improvement initially. This is the essential feature of the log-log law—the slowing down of the learning rate. Hence learning in complex systems will tend to be approximately linear in log-log space.

The great virtue of this argument, or some refinement of it, is that it would explain the ubiquity, even unto the industrial production functions.

We do not know how to examine this law in full generality. However, restriction to a subclass of learning functions, if the subclass is rich enough, can shed some useful light on the issue, for the argument should hold for the subclass as well.

The complete definition of a mixture model requires both the specification of a class of learning functions and a scheme by which they are aggregated. A natural class of learning functions is the exponential functions. They form a rich enough class (a three-parameter family of α, A, and B). They also are as good a candidate as any for primitive learning functions. We can place sufficient restriction on the means of aggregation if we assume that performance consists of the *serial* execution of subtasks. This places us within the class of additive systems, that is, where each component adds its contribution to the total performance.[7] The result is that T is a weighted sum of exponentials:

$$T = \sum_i W_i e^{-\mu_i N} \tag{37}$$

Figure 1.25 shows a plot in log-log space of a 40-term sum with weights (the W's) and rates (the μ's) selected at random ($0 < W_i < 5$ and $0 < \mu_i < .1$). One achieves a reasonable approximation to a straight line over much of the range, though it is a little wavy.

Mixtures of this type have one primary source of variation: the set of weights $\{W_i\}$. The plausibility of mixture models as a source for power laws can best be evaluated by determining the classes of functions that are generated under reasonable assumptions for $\{W_i\}$. If the result is always a power law, then mixture models are strongly implicated. On the other hand, if any function can be generated with equal facility, mixtures would be of little use as an explanation for the ubiquity of power laws.

Sums of exponentials do provide a sufficient ensemble of functions to compose (essentially) any function desired. A convenient way to see this is to go over to the continuous case:

$$T(N) = \int_0^\infty W(\mu) e^{-\mu N}\, d\mu \tag{38}$$

On the one hand, this simply expresses the continuous analog of a sum of exponentials: the exponential for every μ is represented, each with its own weight, $W(\mu)$. On the other hand, this will instantly be recognized (at least by engineers and mathematicians) as the Laplace transform of the function W (Churchill, 1972). The significance of this is that we know that for any function $T(N)$ there is a function $W(\mu)$ that produces it.[8] Thus, by choosing appropriate weights, any total learning function whatsoever can be obtained.

[7] Simple additive combination is not the only way to put learning mechanisms together. Clayton Lewis (no date) explored the notion of series-parallel combinations of exponential learning mechanisms. The results were unclear, sometimes looking log-log, sometimes looking more like an exponential, sometimes wandering. He arrived (1980) at the position that another source of constraint or uniformity is needed.

[8] T must be mathematically well behaved in certain ways to be so represented, but this is of no consequence in the present context.

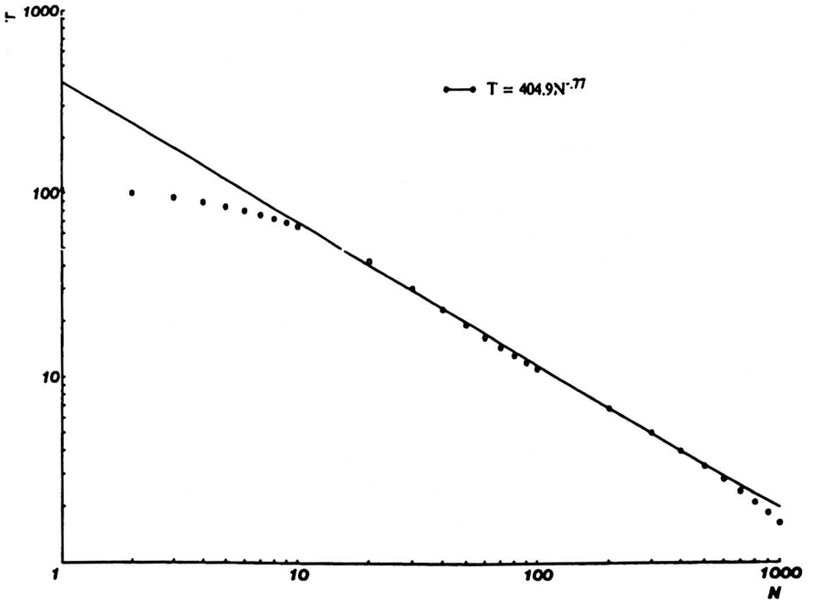

FIG. 1.25. A forty term additive exponential mixture (log-log coordinates). The weights ($0 < W_i < 5$) and exponents ($0 < \mu_i < .1$) were selected at random.

We can of course choose weights to make T a power law, as in Equation 4, with α and B. Consulting any standard table of Laplace transforms shows

$$W(\mu) = \left[\frac{B}{\Gamma(\alpha)} \right] \mu^{-(1-\alpha)} \tag{39}$$

That is,

$$T(N) = BN^{-\alpha} = \int_0^\infty \left[\frac{B}{\Gamma(\alpha)} \right] \mu^{-(1-\alpha)} e^{-\mu N}\, d\mu \tag{40}$$

The component exponentials correspond to learning at all rates, indefinitely fast (large μ) to indefinitely slow (small μ). Because $(1 - \alpha) \geq 0$, the weight W becomes very small for fast learning and very large for slow learning. Without a justification for this particular distribution of weights, it would seem implausible that mixtures of learning components would always lead to power laws.

However, we can turn the argument around and obtain a positive result. One distribution of weights for which there is a natural justification is the rectangular (i.e., all component processes have the same weight, at least stochastically). This is especially true in the present approximation, where a random distribution of weights would be taken to be rectangular. As seen from Equations 39 and 40, this

corresponds to $(1 - \alpha) = 0$, which yields $\alpha = 1$. The resulting law is the hyperbolic.

It is beyond the bounds of this chapter to inquire how closely random weighting functions can be approximated by the mean. Within our limits, it appears that a mixture of exponentials yields a special case of the power law, namely, the hyperbolic. Put together with the results of the data-fit analysis, which showed that hyperbolics were a reasonable candidate descriptive curve, this adds up to a significant observation (it can hardly be distinguished as a "result").

Real mixtures can only strive to approximate the distribution of exponentials that the use of rectangular weights implies. They must fall short because there can only be a finite number of components. The initial portion of Fig. 1.25 is flattened because of the lack of terms in the mixture that decay quickly enough to affect that portion. We restricted the fastest term to have a μ less than .1, but there must always be a maximum μ. Regions of the curve that are affected by only a few terms will look highly exponential, leading to a roller coaster effect where two such regions meet [e.g., for N in the region (10,200) in Fig. 1.25]. In regions where only one term is relevant, the curve is an exponential. This must always occur at least in the tail of the curve, where only the slowest term in the mixture is still active.

The amount of deviation within a region of the curve is thus determined by the number of terms affecting that region. Linearity over a wide range requires a large number of terms in the mixture.

Stochastic Selection

The work in stochastic modeling generated a large range of models, well beyond what we can review. However, a few of the models are particularly relevant to this work.

Crossman's Model

Twenty years ago, Crossman (1959), in an effort similar in spirit to the present one, wrote a paper reviewing much data on practice. He proposed a general model based on an improving process of selecting methods from a fixed population of methods with fixed durations, $\{t_i\}$. Improvement occurs, because each method is selected according to a probability and these probabilities are adjusted on the basis of experience. Namely, the change in probability is proportional to the difference between the mean time, $T(N)$, and the actual time of the selected method, t_i:

$$\delta p_i = -k[t_i - T(N)] \tag{41}$$

By assuming that the entire probability vector shifts at each trial according to its expected adjustment (i.e., as if all methods were tried each trial, each with frequency p_i), the expected shift for the mean time can be expressed as

$$T(N + 1) = T(N) - k\text{Var}(N) \tag{42}$$

where $\text{Var}(N)$ is the variance of the $\{t\}$ on cycle N.

In general, the time course cannot be calculated without knowing the actual distribution of the t_i, for the following relationships hold for this model [$M_j(N)$ is the jth moment of the $\{t_i\}$ on cycle N]:

$$T(N) = M_1(N) \tag{43}$$

$$\text{Var}(N) = M_2(N) - [M_1(N)]^2 \tag{44}$$

$$M_j(N + 1) = [1 + kM_1(N)]M_j(N) - kM_{j+1}(N) \tag{45}$$

Thus, as N increases, higher moments of the initial distribution are needed to compute $\text{Var}(N)$. Crossman assumed an (somewhat arbitrary) example distribution and examined the resulting curve numerically. In log-log space it plotted as a sigmoid with a large straight section, somewhat in the manner of Fig. 1.15. He concluded that it was a satisfactory form of model, though clearly needing more development.

Unfortunately, the model rests very heavily on the way it uses its expected value assumptions. As seen from Equation 41, nothing prevents p_i from moving outside the [0, 1] interval, thereby violating the basic property of being a probability. Indeed, if the ith method is selected often enough, it must move outside. Crossman avoids the unavoidable by making the change really be $p_i\delta p_i$, the expected change. Even this modification is not sufficient to guarantee that p_i remains in the range of [0, 1]. If k is greater than $1/(t_{\max} - t_{\min})$, then it is possible for $\delta p_{\min}$ to be less than -1. An additional assumption about the legal values of k could of course be added to handle this problem.

We have expounded Crossman's model at some length, because not only is it the one existing attempt to deal with the power-law data, but it is often referred to as a viable explanation of this law.

The Accumulator and Replacement Models

Among the basic stochastic learning models two broad classes are often distinguished, depending on whether correct responses replace incorrect ones—called *replacement* models—or whether correct responses are simply added to the total pool, thus gradually swamping out the incorrect ones—called *accumulator* models. A presentation of these two models is given in Restle and Greeno (1970).

The replacement models yield exponential functions (when expressed in terms of rate of generation of correct responses). It is worth taking a look at an accumulator model, as it will provide another model that yields the hyperbolic. Restle and Greeno show that the proportion of correct responses in the pool at trial N (P_N) is given by (the interpretations of the other parameters are not important for our purposes)

$$P_N = \frac{b + \theta a(N - 1)}{1 + \theta(N - 1)} \tag{46}$$

To get this in terms of time, we can assume that the time to generate a response is inversely proportional to the rate of generation of correct responses. Thus $T(N)$ would be the inverse of Equation 46:

$$T(N) = \frac{1 + \theta(N - 1)}{b + \theta a(N - 1)} \tag{47}$$

With a little rearrangement, this becomes:

$$T(N) = \frac{1}{a} + \frac{(a - b)/\theta a^2}{N + [b/\theta a - 1]} \tag{48}$$

This is the equation for a general hyperbolic function, with $A = 1/a$, $B = (a - b)/\theta a^2$, and $E = b/\theta a - 1$.

Exhaustion of Exponential Learning

The notion of exhaustion comes from examining Equation 11. A power law is like an exponential in which the exponent (α) does not remain constant over trials. In fact, α decreases as $1/N$. An exhaustion model would postulate that this decrease stems from the diminishment of some necessary portion of the learning process. Many different exhaustion models can be developed according to what is being diminished. We have concentrated our efforts on one variety of exhaustion model; what we call the *chunking model of learning*. Before we examine it in detail, it is useful to look briefly at the range of possible exhaustion models. In the descriptions that follow it is assumed that the learner uses some *method* for the performance of the task on which he is working. Learning consists of finding and incorporating *improvements* to the current method.

- *Improvements harder to find (search exhaustion):* Improvements may not always be right at hand. It would then be necessary to search for improvements that can be made in the method being used. Each time one is found, it would result in the time (T) decreasing by some constant factor (α), just as in exponential learning. As improvements are found and applied, the space of unused improvements becomes sparser, decreasing the rate at which new improvements can be found. The effective rate of learning would thus be slowed.
- *Less time for improvement (time exhaustion):* If learning is exponential in time (rather than in trials), then as the trials get shorter, there is less time for improvement on each trial. We saw earlier that an exponential in time yields a hyperbolic in trials.

One long standing view is that learning consists of transforming a deliberate, conscious and resource limited process into an automatic, unconscious and resource independent one. One image of this in mechanism is that learning consists of a transformation from a serial to a parallel processing structure. The amount of processing required remains constant. Only the elapsed time until completion decreases. Exhaustion occurs if it is assumed that learning is proportional to the amount of time available (T)—the usual exponential assumption. As the amount of process that is packed into a fixed time slice increases, the amount of learning per unit of process would have to decrease. A simple version of this model that we have developed yields the hyperbolic.

- *Improvements less effective (effectiveness exhaustion):* Improvements used later in learning, may prove to be less effective than the same improvements used earlier.
- *Improvements less applicable (applicability exhaustion):* Improvements may vary from being general purpose to being highly specialized. General-purpose improvements are always applicable, while special purpose ones may only be applicable under highly constrained conditions. In order to specify a model of this type fully, an assumption must be made as to the order in which the improvements are incorporated into the method. If they are used in order of decreasing applicability, then learning would slow down even if the improvements are equal in effectiveness (when they are applicable). The theory we present now is a version of this case.

The Chunking Theory of Learning

We take as central to our model a theme that has been a mainstay of information-processing psychology since Miller's famous 1956 paper.

> *The Chunking Hypothesis:* A human acquires and organizes knowledge of the environment by forming and storing expressions, called *chunks*, which are structured collections of the chunks existing at the time of learning.

This brief statement glosses over things not central to our purpose, for example: (1) the nature of the primitive chunks; (2) the internal representation of chunks as collections of symbols for chunks, rather than the chunks themselves; and (3) distinctions, if any, between perceptual chunks, internal-processing chunks, and motor chunks. Other aspects, such as the size and composition of chunks, require further specification.

Consider Seibel's task (Seibel, 1963), to make matters concrete. There are ten lights $L_1, \ldots, L_{10}$, which define perceptual events of a light being off ($-$) or on ($+$). Originally, the only chunks available are the individual lights and the states of *off* and *on*. If we define the notion of the *span* of a chunk as the number of primitive elements that it contains, then these are chunks with a span of one. Clearly they are built up from still more primitive features, relations, etc., but they can be taken as the primitives from the point of view of Seibel's task. Gradually, with learning, chunks will form: first chunks such as $(L_1 +)$, which we might also write as L_1^+; then chunks such as $(L_3^+ L_4^+)$ or $(L_1^- L_{10}^-)$; then still

higher chunks such as $(L_2^+(L_3^-L_4^-))$, and so on. The chunks need not just be of perceived lights; they could be of responses $(R_5^+R_6^+)$ (the + meaning to press the button) or even of mixed character, $(L_3^+R_3^+)$ or $((L_7^+L_8^-)(R_7^+R_8^-))$. These chunks are of increasing span; for example, the span of the last mentioned chunk, $((L_7^+L_8^-)\ (R_7^+R_8^-))$, is eight of the primitive chunks such as L_7, +, L_8. Chunks thus hold information about the *patterns* in the environment and in the subject's relation to the environment.

The chunking assumption only defines a unit of structure and declares it central. To create a learning system, we must tie down how this structure couples to: (1) the performance of the task; (2) the structure of the task environment; and (3) the process of learning new information about the task environment. These lead to three corresponding general assumptions:

- *Performance assumption:* The performance program of the system is coded in terms of high-level chunks, with the time to process a chunk being less than the time to process its constituent chunks.
- *Task structure assumption:* The probability of recurrence of an environmental pattern decreases as the pattern size increases.
- *Learning assumption:* Chunks are learned at a constant time rate on average from the relevant patterns of stimuli and responses that occur in the specific environments experienced.

On Performance. If having chunks does not permit the system to perform more quickly, then one major reason for their existence vanishes (though there might be other reasons). How high-level aggregate chunks enter into performance programs is actually somewhat problematical. For instance, computers gain no performance advantage from the subroutine hierarchy (an example of multilevel chunking); it is completely unwound down to the lowest level machine operations on every execution.

In Seibel's task the performance program can be related directly to the chunks that exist. If only the lowest chunks are available, then it might take the processing of five chunks for each light:

$$\begin{array}{l}((L_x\ +)(R_y\ +))\\ \quad L_i\ +\ \ R_i\ +\end{array}$$

The top chunk is the rule derived from the instructions for general lights (L_x) and responses (R_y); it is used to interpret each of the four primitive chunks of information about the task, one after the other. If, on the other hand, more complete chunks are available, such as $(L_i\ +)$, then this part can be done in a single step, and so on for more aggregate chunks. Aggregation, of course, takes place not just within a light, but across lights. Thus, a lowest-level performance program would take something like 5 steps per light times 10 lights = 50 steps. At the other extreme, the highest-level program would take only a single step,

using many mammoth chunks, such as the one below of span 40, to cover all the cases.

$$((((L_1^+R_1^+)(L_2^-R_2^-))(L_3^+R_3^+))\ ((L_4^+R_4^+)(L_5^+R_5^+))\ ((L_6^+R_6^+)((L_7^-R_7^-)(L_8^-R_8^-))$$
$$((L_9^-R_9^-)(L_{10}^+R_{10}^+)))) \qquad (49)$$

Most programs would be composed of chunks of some intermediate span. Our example chunks have used stimulus adjacency and stimulus–response connection as the principles on which to chunk. Lots of others are possible (e.g., symmetry of position). Likewise, wrong connections are possible as well as correct ones.

On the Structure of the Task Environment. Task environments can be thought of as being composed from a set of elements that can vary with respect to attributes, locations, relations to other elements, etc. Seibel's task is a good example of such a task environment once chunking has reached beyond the most primitive level (the lights, on, off, etc.). Observe that (thinking only about the lights) there is a set of elements (the 10 lights) each of which has an attribute for the state of the light (*on* or *off*). On each trial the subject is exposed to a single *concrete environment* out of the *ensemble* of concrete environments that make up the *task environment*. A subject in Seibel's experiment would see the 10 lights in one particular state on each trial. The trial sequence provides the sample of concrete environments actually experienced.

Figure 1.26 shows a four-light version of Seibel's task environment. At the left are the primitive chunks, the lights, which can be either on or off. Proceeding toward the right yields higher level chunks made by combining lower-level ones. At the far right are the top-level chunks. Each top-level chunk spans one concrete environment (consisting of each light in one particular state). The bold lines outline one concrete environment out of the ensemble that makes up the task environment. One important point to notice is that the branchiness of the task environment (increasing toward the right) is in the opposite direction from that of the tree for a single concrete environment (increasing toward the left). As the chunks increase in span, there are more of them in the task environment but fewer in any one concrete environment.

Task environments such as Seibel's present the learner with a combinatorial number of possible patterns. There are only 2 patterns of 1 light (on and off), but 4 patterns of 2 lights, 8 patterns of 3 lights, and so on, up to 1024 patterns of 10 lights. Inherently, many more possibilities for patterns of elements exist than for the elements themselves. Correspondingly, there are many more possibilities for chunks that encode larger patterns than smaller ones. If each of the elements can take on any of b different values (the *branchiness* of the task environment), then for every set of s elements there would be b^s possible patterns. Different task environments will have constraints that limit what new combinations can in fact occur; not all elements are or can be chunked with each other. The basic combinatorial nature of most task environments, combined with these constraints, will determine what can be called the *cardinality* of the task environment,

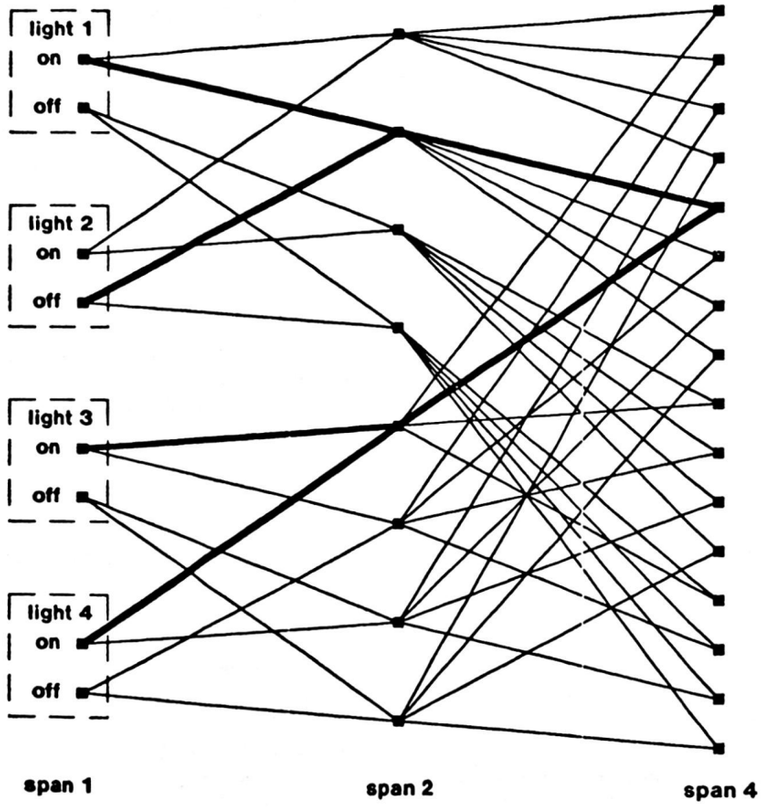

FIG. 1.26. Seibel's task environment for four lights. At the left are the two primitive chunks for each light (for the *on* and *off* states) and at the right are the *top-level* chunks.

namely, the number of patterns that can actually occur of each different span. This cardinality (whether exponential, power law, etc.) will have a great deal to do with the form of the final learning curve.

The *task structure assumption* follows directly from this structure of the task environment. There are more of the larger patterns, but each one appears in fewer concrete environments. Indeed, at the topmost level, the entire concrete environment at a trial can be encoded in a single chunk, as in Example 49. Chunks of this type appear in only one concrete environment each, whereas a chunk that only contains a single light and its state would appear in many concrete environments. The multiplicity of patterns (chunks) depends on there being an entire ensemble of possible concrete environments. In any particular concrete environment only a small number of the possible chunks occur.

On Learning by Experience. This assumption starts from the view that the human is a time-independent processing mechanism. It processes information the same way one hour as the next, one day as the next—as a function of stored knowledge and learned procedures but not of time per se. In short, there is no built-in historical clock. Thus, there exists a basic constant rate of chunk acquisition (with respect to time, not trials). This same view underlies the appeal of the *total time hypothesis* of verbal learning (Cooper & Pantle, 1967).

Not all chunks learned need be relevant to the task at hand. The assumption that learning is by experience says the subject is picking up relevant chunks while performing in a concrete environment. This is consonant with theories that have learning occurring automatically from the chunks that are built in *working memory* (involving both the stimuli and the subject's own responses). When the subject is attending to the task, working memory is full of task-related chunks, and relevant learning occurs.

In our example, given L_1 and + perceived by the subject, the chunk $(L_1\ +)$ could be built, but not the chunk $(L_1\ -)$. Also, it would take the same length of time to build the first-level chunk as to build $(((L_1\ +)\ (R_1\ +))\ ((L_2\ -)\ (R_2\ -)))$ given that the constituent chunks, $((L_1\ +)\ (R_1\ +))$ and $((L_2\ -)\ (R_2\ -))$, were available in the subject (i.e., had already been learned) and were being perceived in the environment.

These three assumptions, though still general, provide a basis on which specific learning models can be built. In this chapter we only present the simplest form of this model so that the basic mechanisms can be clearly seen. Various limiting conditions and the like may appear a little strained in this simple version.

A Simple Version

For the theory to be specific, we need to determine T as a function of N. One way to do this is to define the differential learning law, dT/dN. Corresponding to the previous assumptions, we introduce the following variables:

C = the total number of chunks learned at any time.

s = the span to which the subject has chunked.

In terms of these variables, we can compose dT/dN as follows:

$$\frac{dT}{dN} = \left(\frac{dT}{ds}\right)\left(\frac{ds}{dC}\right)\left(\frac{dC}{dN}\right) \tag{50}$$

The first term, dT/ds, expresses how performance time (T) changes with the chunk span. In a simple form of our performance assumption, the time to perform the task will simply be proportional to the number of high-level chunks it takes to describe the task (at the time of the performance). Let P be the number of chunks involved in the performance initially (and take the unit of time to be the time to process one chunk, so as to avoid an arbitrary constant). Then, if chunking has proceeded to a span of s, each top-level chunk spans s initial chunks.

Thus, the number of top-level chunks that are required to span the performance is P/s and we find for the performance time,

$$T = \frac{P}{s} \tag{51}$$

$$\frac{dT}{ds} = \frac{-P}{s^2} = \frac{-T^2}{P} \tag{52}$$

If this holds for unlimited values of s, it implies that P is infinitely divisible and that T can be driven to zero. We just accept such simplifications for the purposes of this model. Given this simplification, however, we cannot expect to find an asymptote parameter (A) in this version.

The second term of Equation 50, ds/dC, expresses how fast the span of the chunks increases as the subject accumulates more chunks. It depends on how many chunks of each span are needed to describe the task environment. According to the assumption about the structure of the task environment, new chunks will be formed to encompass larger patterns in the environment. If a chunk covers a pattern of some set of elements, then it will be relevant to connect it with a certain number of additional elements in the environment to form the next higher level of chunk. We postpone until later the quantification of this process. For now we can just talk in terms of $C_{te}(s)$, the number of chunks needed to cover all patterns of s elements or less in the task environment.

We need to relate $C_{te}(s)$ to $C(N)$, the number of chunks that the subject has at a given trial. By the nature of how chunks are learned, low-level chunks must be acquired before higher-level chunks; that is, chunks are learned from the bottom up. If C chunks have been learned, they will constitute a pyramid up from the bottom. By making the further simplifying assumption that the pyramid is acquired layer by layer (i.e., if the subject has learned C chunks, these will consist of all the chunks provided by the environment from the elementary chunks up to some span), we can equate C and C_{te}.[9] Hence we find

$$C = C_{te} \tag{53}$$

$$\frac{dC}{ds} = C'_{te}(s), \text{ writing } C'_{te}(s) \text{ for } \frac{dC_{te}(s)}{ds} \text{ for clarity} \tag{54}$$

$$\frac{ds}{dC} = \frac{1}{C'_{te}(s)} \tag{55}$$

The final term of Equation 50 follows directly from the assumptions on learning: that the number of chunks learned per unit time is a constant, say λ chunks:

[9] We are glossing over three complications to this picture: (1) M elements can be covered by chunks of span s in a number of ways, depending on how the M elements are partitioned into groups of size s; (2) M elements can be covered by a number of different chunks of span M that vary in internal structure; and (3) many patterns in the environment are totally irrelevant to performance on the task.

$$\frac{dC}{dt} = \lambda \tag{56}$$

Therefore by Equation 22, which relates time to trials,

$$\frac{dC}{dN} = \left(\frac{dC}{dt}\right)\left(\frac{dt}{dN}\right) = \lambda T \tag{57}$$

We now have assembled all the components of Equation 50:

$$\frac{dT}{dN} = \left(\frac{-T^2}{P}\right)\left[\frac{1}{C'_{te}(s)}\right](\lambda T) \tag{58}$$

$$= \frac{-\lambda}{P}\left[\frac{T^3}{C'_{te}(s)}\right] \tag{59}$$

We can see in what sense this is an exhaustion model. The subject continues to learn at a constant rate and chunks remain equally potent in terms of what they do to the performance programs in which they occur. However, the chance that a chunk will be used becomes increasingly rare. It becomes rarer, actually, because of the increased span of the chunk, which makes it ever more specialized, thus occurring in ever fewer environments. However, this turns out to be correlated with time, because general (i.e., low level) chunks are learned first and specialized chunks are learned later.

A Combinatorial Task Environment

To complete the definition of the chunking model it will be necessary to be more specific about $C_{te}(s)$, the cardinality of the task environment, which expresses how fast the number of patterns increase as their span increases. One possibility is to start from the basic combinatorial structure described under the task structure assumption. Suppose there are M elements in the task environment, each with b possible values. We need to know how many chunks of span s it takes to cover the task environment. One way to do this is to partition the task environment into M/s groups of s elements. It will take b^s chunks of span s to cover each group and so $(M/s)b^s$ chunks to cover the whole task environment. We thus find

$$C_{te}(s) = \sum_{i=1}^{s}\left(\frac{M}{i}\right)b^i \tag{60}$$

This summation does not have a closed form solution. We can however derive $C'_{te}(s)$ directly from the summation in the same manner that dt/dN is obtained in Equation 22.

$$C'_{te}(s) = \left(\frac{M}{s}\right)b^s = \left(\frac{M}{s}\right)e^{\beta s} \tag{61}$$

where $\beta = \log(b)$.

Substituting $C'_{te}(s)$ into Equation 59, we find:

$$\frac{dT}{dN} = -\left(\frac{\lambda}{PM}\right) T^3 s e^{-\beta s} \tag{62}$$

We can eliminate s by noticing that $s = P/T$ (from Equation 51):

$$\frac{dT}{dN} = -\left(\frac{\lambda}{M}\right) T^2 e^{-\beta P/T} \tag{63}$$

By suitable rearrangement and integration, the final form of the learning curve is obtained:

$$T^{-2} e^{\beta P/T}\, dT = -\left(\frac{\lambda}{M}\right) dN \tag{64}$$

$$\int T^{-2} e^{\beta P/T}\, dT = -\left(\frac{\lambda}{M}\right) \int dN \tag{65}$$

$$(\beta P)^{-1} e^{\beta P/T} = \left(\frac{\lambda}{M}\right)(N + E) \tag{66}$$

where E comes from the integration constant.

$$T = \frac{\beta P}{\log(\lambda \beta P/M) + \log(N + E)} \tag{67}$$

Though this is not a power law, it does resemble one when plotted in log–log coordinates. Figure 1.27 shows such a learning curve with parameters of $b = 2$, $P = 50$, $\lambda = 1$, $M = 20$, and $E = 10$. The reason for this linearity can best be seen by looking at dT/dN. Substituting for $1/T$ in the exponent of Equation 63 yields

$$\frac{dT}{dN} = -\left[\frac{(\beta P)^{-1}}{N + E}\right] T^2 \tag{68}$$

$$= -\left[\frac{\alpha}{N + E}\right] T \tag{69}$$

where $\alpha = T/\beta P$.

The function thus behaves like a power law with a slowly decreasing α. In log–log space the decreasing α is difficult to distinguish from the presence of an asymptote.

The Power Law Chunking Model

Instantiations of the chunking model can be generated for various types of task environment that a learner may have to deal with. There is no space here to examine possible task environments systematically. An alternative is to determine what type of task environment leads the chunking model to predict power-law learning. From Equation 8 we know that one form for the differential of a power law is

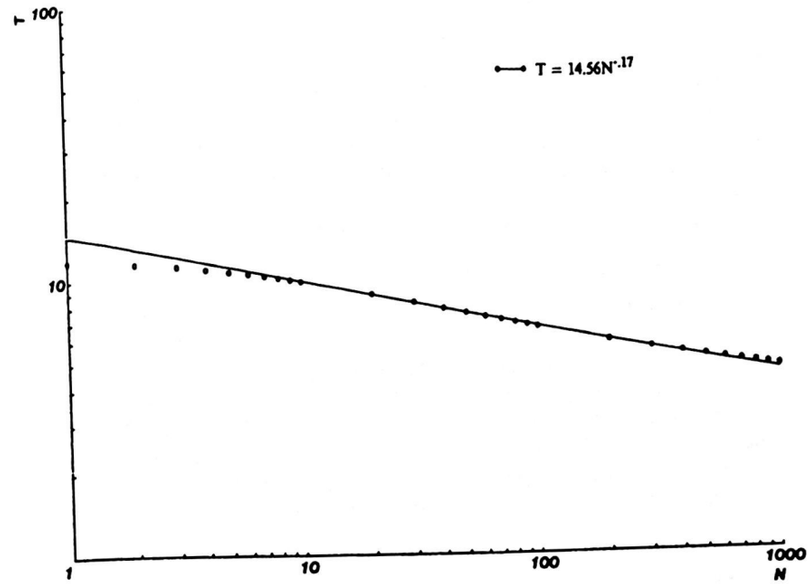

FIG. 1.27. The learning curve for the *chunking model* in a combinatorial task environment (log–log coordinates). The parameter values are: $b = 2$, $P = 50$, $\lambda = 1$, $M = 20$, and $E = 10$.

$$\frac{dT}{dN} = -\alpha B^{-1/b} T^{1+1/b} \tag{70}$$

Combining this with Equation 59 yields

$$-\alpha B^{-1/b} T^{1+1/b} = -\left(\frac{\lambda}{P}\right)\left[\frac{T^3}{C'_{te}(s)}\right] \tag{71}$$

We want $C_{te}(s)$, so first solving for $C'_{te}(s)$

$$C'_{te}(s) = \left(\frac{\lambda B^{1/b}}{P\alpha}\right) T^{2-1/b} \tag{72}$$

$$= \left(\frac{\lambda B^{1/b}}{P\alpha}\right)\left(\frac{P}{s}\right)^{2-1/b} \tag{73}$$

$$= \left(\frac{\lambda B^{1/b} P^{1-1/b}}{\alpha}\right) s^{1/b-2} \tag{74}$$

Now we can find $C_{te}(s)$ by integrating $C'_{te}(s)$ with respect to s.

$$C_{tr}(s) = \left[\frac{\lambda B^{1/n} p^{1-1/n}}{1-\alpha}\right] s^{1/n-1} \qquad (75)$$

Though it is somewhat obscured by the complex initial constant, this is a power law in s. Power-law learning thus implies a power-law environment. An important, and indeed pleasing, feature of the chunking model is this connection between the structure of the task environment and the learning behavior of the subject. The richer the task environment (i.e., the ensemble of environments with which the subject must potentially cope) the more difficult his learning.

Relation to Existing Work on Chunking

An important aspect of the chunking model of learning is the amount of power it gets by making connection with a wide body of existing psychological work. For example, the pervasiveness of the phenomenon of chunking amply accounts for the ubiquity of log-log learning. We have been able to develop the primary assumptions of the model from this work without the necessity of pulling an arbitrary "natural" learning curve out of the air.

Much of the existing work on chunking has focused on showing that chunks are the structures of memory and operate in behavior in various ways (Bower & Winzenz, 1969; Johnson, 1972). It is consonant with the present model but does not make interesting contact with it. However, the work on chess perception (Chase & Simon, 1973; DeGroot, 1965) bears directly on the present model. The basic phenomenon investigated there was the differential short-term memory for meaningful chess positions with expertness. Novices are able to recall only a few pieces of a complex middle-game position after a 5-second exposure, whereas masters can recall most of the pieces.

A well-articulated theory has evolved to explain this striking phenomenon. The theory is an elaboration of the basic assumptions about chunking. The master has acquired an immense memory for chess positions, organized as a collection of chunks. His ability for immediate perception and short-term memory of chess positions depends directly on how many chunks are used to encode a position. Estimates of the number of chunks available to the master are of the order of 50,000, based on extrapolation of a simulation program (Simon & Gilmartin, 1973) that fits novice- and expert-level players. By implication, master players must spend an immense amount of time with the game, in order to acquire the large number of chunks; this seems to be well supported by historical data.

The chunking model of learning presented here for the power law is essentially the same as the chess perception model. The present model has been elaborated quantitatively for learning data, whereas the chess perception data had the products of learning to work with. The explanation for why the number of perceptual chess chunks is so large lies in the combinatorial complexity of chess positions. High-level chess chunks encode large subpatterns of pieces on the board; they are the necessary means for rapid perception. But the actual configurations to which they apply do not show up often. Thus to gain coverage of the population of chess positions requires acquisition of immense numbers of high-level chunks. This is precisely the notion of environmental exhaustion that is the key mechanism of the present model.

One would expect from this that the time course of chess skill would also follow the power law, if one would take the trouble to measure it. Indeed, the data on the *Stair* game of solitaire in Fig. 1.10 can be taken as a reasonable analog of the chess game.

CONCLUSION

If we may, let us start this conclusion by recounting our personal odyssey in this research. We started out, simply enough, intrigued by a great quantitative regularity that seemed to be of immense importance (and of consequence for an applied quantitative psychology), well known, yet seemingly ignored in cognitive psychology. We saw the law as tied to skill and hence relevant to the modern work in automatization. The commitment to write this chapter was the goad to serious research. When we started, our theoretical stance was neutral—we just wanted to find out what the law could tell us. Through the fall of 1979, in searching for explanations, we became convinced that plausible substantive theories of power laws were hard to find, though it seemed relatively easy to obtain an exponent of -1 (i.e., hyperbolics). In November, discovering the chunking model (by looking for forms of exhaustion, in fact), we became convinced that it was the right theory (at least A. N. did) and that lack of good alternative theories helped to make the case. The chunking model also implied that the power law was not restricted to perceptual-motor skills but should apply much more generally. This led to our demonstration experiment on *Stair*, which showed a genuine problem-solving task to be log-log linear. At the same time, in conversations with John Anderson, additional data emerged from the work of his group (Figs. 1.7 and 1.9) that bolstered this.

This picture seemed reasonably satisfactory, though the existence of log-log linear industrial learning curves (Fig. 1.12) nagged a bit, as did the persistence of some of our colleagues in believing in the argument of mixtures. However, as we proceeded to write the chapter, additional work kept emerging from the literature, including especially the work by Mazur and Hastie (1978), that raised substantial doubts that the power law was the right empirical description of the data. The resulting investigation has brought us to the present chapter.

The picture that emerges is somewhat complex, though we believe at the moment that this complexity is in the phenomena, and not just in our heads as a reflection of only a momentary understanding. We summarize this picture below, starting with the data and progressing through theoretical considerations.

1. The empirical curves do not fit the exponential family. Their tails are genuinely slower than exponential learning and this shape deviation does not disappear with variation of asymptote.

2. The data do satisfactorily fit the family of generalized power functions (which includes the hyperbolic subfamily). There is little shape variance remaining in the existing data to justify looking for other empirical families.

In particular, there is no reason to treat apparent systematic deviations, such as occur in Snoddy's or Seibel's data in log-log space (Figs. 1.1, 1.6), as due to special causes, distinct from their description as a generalized power function.

3. The data do not fit the simple power law (i.e., without asymptote or variable starting point). There are systematic shape deviations in log-log space (the space that linearizes the simple power law), which disappear completely under the general power law.

4. We were unable to confirm either whether the data fit within the hyperbolic subfamily or actually requires the general power family. This is so despite the multitude of existing data sets, some with extremely lengthy data series (some of it as extensive as any data in psychology).

5. The major phenomenon is the ubiquity of the learning data (i.e., its common description by a single family of empirical curves). We extended the scope to all types of cognitive behavior, not just perceptual-motor skill.

However, we restricted our view to performance time as the measure of performance, though learning curves measured on other criterion also yield similar curves. Also, we restricted our view to clear situations of individual learning, though some social (i.e., industrial) situations yield similar curves. Our restriction was dictated purely by the momentary need to bound the research effort.

6. Psychological models that yield the power law with arbitrary rate (α) are difficult to find. (Positive asymptotes and arbitrary starting points are, of course. immediately plausible, indeed, unavoidable.)

7. Models that yield the hyperbolic law arise easily and naturally from many sources—simple accumulation assumptions, parallelism, mixtures of exponentials, etc.

8. The various models are not mutually exclusive but provide an array of sources of the power law. Several hyperbolic mechanisms could coexist in the same learner. Independent of these, if the humans learn by creating and storing chunks, as there is evidence they do, then the environmental-exhaustion effect would also operate to produce power-law learning, independent of whether there were other effects such as mixing to produce hyperbolic learning curves.

9. A maintainable option is that the entire phenomenon is due to exponential component learning yielding an effective hyperbolic law through mixing.

This would cover not only the data dealt with here but probably also the data with other criteria and the data from industrial processes.

However, the exponential learning of the component learners remains unaccounted for.

10. The chunking model provides a theory of the phenomena that offers qualitatively satisfactory explanations for the major phenomena.

However, some of the phenomena, such as the industrial processes, probably need to be assigned to mixing. Parsimony freaks probably will not like this.

The theory is pleasantly consistent with the existing general theory of information processing and avoids making any a priori assumptions.

Though power laws are not predicted for all task environments, the learning curves do closely approximate power laws.

ACKNOWLEDGMENTS

This research was sponsored in part by the Office of Naval Research under contract N00014-76-0874 and in part by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 3597, monitored by the Air Force Avionics Laboratory under contract F33615-78-C-1551.

The views and conclusions in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research, the Defense Advanced Research Projects Agency, or the U.S. Government.

REFERENCES

Anderson, J. Private communication, 1980.

Bower, G. H., & Winzenz, D. Group structure, coding, and memory for digit series. *Experimental Psychology Monograph*, 1969, *80*, 1-17 (May, Pt. 2).

Calfee, R. C. *Human experimental psychology*. New York: Holt, Rinehart & Winston, 1975.

Card, S. K., English, W. K., & Burr, B. Evaluation of mouse, rate controlled isometric joystick, step keys, and text keys for text selection on a CRT. *Ergonomics*, 1978, *21*, 601-613.

Card, S. K., Moran, T. P., & Newell, A. Computer text editing: An information-processing analysis of a routine cognitive skill. *Cognitive Psychology*, 1980, *12*(1), 32-74. (a)

Card, S. K., Moran, T. P., & Newell, A. The keystroke model for user performance time with interactive systems. *Communications of the ACM*, 1980, *23*. (In press; available as SSL-79-1, Xerox PARC). (b)

Chase, W. G., & Simon, H. A. Perception in chess. *Cognitive Psychology*, 1973, *4*, 55-81.

Churchill, R. V. *Operational mathematics*. New York: McGraw-Hill, 1972.

Cooper, E. H., & Pantle, A. J. The total-time hypothesis in verbal learning. *Psychological Bulletin*, 1967, *68*, 221-234.

Crossman, E. R. F. W. A theory of the acquisition of speed-skill. *Ergonomics*, 1959, *2*, 153-166.

Crowder, R. G. *Principles of learning and memory*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1976.

deGroot, A. D. *Thought and choice in chess*. The Hague: Mouton, 1965.

DeJong, R. J. The effects of increasing skill on cycle-time and its consequences for time-standards. *Ergonomics*, 1957, *1*, 51-60.

Fitts, P. M. The information capacity of the human motor system in controlling amplitude of movement. *Journal of Experimental Psychology*, 1954, *47*, 381-391.

Fitts, P. M. Perceptual-motor skill learning. In A. W. Melton (Ed.), *Categories of human learning*. New York: Academic Press, 1964.

Fitts, P. M., & Posner, M. I. *Human performance*. Belmont, Calif.: Brooks/Cole, 1967.

Guilliksen, H. A rational equation of the learning curve based on Thorndike's law of effect. *Journal of General Psychology*, 1934, *11*, 395-434.

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

Hirsch, W. Z. Manufacturing progress functions. *Review of Economics and Statistics*, 1952, *34*, 143-155.

Hyman, R. Stimulus information as a determinent of reaction time. *Journal of Experimental Psychology*, 1953, *45*, 188-196.

Johnson, N. F. Organization and the concept of a memory code. In A. W. Melton & E. Martin (Eds.), *Coding processes in human memory*, Washington, D.C.: Winston, 1972.

Kintsch, W. *Memory and cognition*. New York: Wiley, 1977.

Kolers, P. A. Memorial consequences of automatized encoding. *Journal of Experimental Psychology: Human Learning and Memory*, 1975, *1*(6), 689-701.

LaBerge, D. Acquisition of automatic processing in perceptual and associative learning. In P. A. M. Rabbitt & S. Dornic (Eds.), *Attention and performance V*, New York: Academic, 1974.

Lewis, C. Speed and practice, undated.

Lewis, C. Private communication, 1980.

Lindsay, P., & Norman, D. *Human information processing: An introduction to psychology* (2nd ed.). New York: Academic, 1977.

Mazur, J., & Hastie, R. Learning as accumulation: A reexamination of the learning curve. *Psychological Bulletin*, 1978, *85*(6), 1256-1274.

Miller, G. A. The magic number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 1956, *63*, 81-97.

Moran, T. P. Compiling cognitive skill, 1980 (AIP Memo 150, Xerox PARC).

Neisser, U., Novick, R., & Lazar, R. Searching for ten targets simultaneously. *Perceptual and Motor Skills*, 1963, *17*, 955-961.

Neves, D. M., & Anderson, J. R. Knowledge compilation: Mechanisms for the automatization of cognitive skills. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition*, Hillsdale, N.J.: Lawrence Erlbaum Associates (in press).

Newell, A. Harpy, production systems and human cognition. In R. Cole (Ed.), *Perception and production of fluent speech*, Hillsdale, N.J.: Lawrence Erlbaum Associates, 1980.

Posner, M. I., & Snyder, C. R. R. Attention and cognitive control. In R. L. Solso (Ed.), *Information processing and cognition*, Hillsdale, N.J.: Lawrence Erlbaum Associates, 1975.

Reisberg, D., Baron, J., & Kemler, D. G. Overcoming Stroop interference: The effects of practice on distractor potency. *Journal of Experimental Psychology: Human Perception and Performance*, 1980, *6*, 140-150.

Restle, F., & Greeno, J. *Introduction to mathematical psychology*. Reading, Mass.: Addison-Wesley, 1970 (chap. 1).

Rigon, C. J. Analysis of progress trends in aircraft production. *Aero Digest*, May 1944, 132-138.

Robertson, G., McCracken, D., & Newell, A. The ZOG approach to man-machine communication. *International Journal of Man-Machine Studies* (in press).

Schneider, W., & Shiffrin, R. M. Controlled and automatic human information processing: I. Detection, search and attention. *Psychological Review*, 1977, *84*, 1-66.

Seibel, R. Discrimination reaction time for a 1,023 alternative task. *Journal of Experimental Psychology*, 1963, *66*, 215-226.

Shiffrin, R. M., & Schneider, W. Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, 1977, *84*, 127-190.

Simon, H. A. On a class of skew distribution functions. *Biometrika*, 1955, *42*, 425-440.

Simon, H. A., & Gilmartin, K. A simulation of memory for chess positions. *Cognitive Psychology*, 1973, *5*, 29-46.

Snoddy, G. S. Learning and stability. *Journal of Applied Psychology*, 1926, *10*, 1-36.

Spelke, E., Hirst, W., & Neisser, U. Skills of divided attention. *Cognition*, 1976, *4*, 215-230.

Stevens, J. C., & Savin, H. B. On the form of learning curves. *Journal of the Experimental Analysis of Behavior*, 1962, *5*(1), 15-18.

Suppes, P., Fletcher, J. D., & Zanotti, M. Models of individual trajectories in computer-assisted instruction for deaf students. *Journal of Educational Psychology*, 1976, *68*, 117-127.

Thurstone, L. L. The learning curve equation. *Psychological Monographs*, 1919, 26(114), 51.

Welford, A. T. *Fundamentals of skill*. London: Methuen, 1968.

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