DRAFT MATERIAL: LIMITED DISTRIBUTION FOR COMMENT.
NOT FOR QUOTATION
File: MSA.MSS
00:17 22 May 1980

# Mechanisms of Skill Acquistion and the Law of Practice

Allen Newell and Paul Rosenbloom

May 1980

**Department of Computer Science**
**Carnegie-Mellon University**
**Pittsburgh, Pennsylvania 15213**

Paper to be given at the 16th Annual Carnegie-Mellon Symposium on Cognition, on *Learning and Cognition*, 21-23 May 1980.

This research was sponsored 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.

# MECHANISMS OF SKILL ACQUISITION AND THE LAW OF PRACTICE[1]

## 1. INTRODUCTION

*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 (p156) 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 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 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 pre-attentive processes, in the earlier work, and the possibility of doing multiple complex tasks simultaneously, in the later.

Recently, especially with the paper 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,

---

[1]This paper relies on the data of many other investigators. We are deeply grateful to those who made available original data: John Anderson, Paul Kolers, David Neves, Patrick Rabbitt, and Richard 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.

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 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 science texts (Calfee, 1975, Crowder, 1976, Kintsch, 1977, Lindsay & Normon, 1977). Likewise, it is not a part of the long history of work on the *learning curve* (Thurstone, 1919, Guilliksen, 1934, Restle & Greeno, 1970), 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.

The aim of this paper 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), 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. Such a general regularity as the log-log law might say something interesting about the basic mechanisms of turing knowledge into action. Third, an unfinished manuscript by Clayton Lewis (Note 2) 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 Section UBIQUITOUS we provide many examples of the log-log law and characterize its universality. In Section 3 we perform some basic finger exercises about the nature of power laws. In Section 4 we investigate questions of curve fitting. In Section 5 we address the possible types of explanations for the law. In Section 6 we develop one approach, which we call the Chunking Theory of Learning. In Section 7 we expand on this theory, applying it to some of the phenomena and exploring, albeit briefly, the implications for the architecture of cognition. Finally, in Section 8, we sum up our results.

---

[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.

# 2. 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 will be presenting 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 seem to be engaged in the task. In Table DATASUMMARY we tabulate several parameters of each of the curves. Their definitions will be given at the points in the paper where the parameters are first used.

## 2.1. 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 SNODDY plots the log of performance on the vertical axis against the log of the trial number for a single subject.

The first important point is:

- The law holds for performance measured as the *time* to achieve a fixed task.
- [SHOULD WE BUILD AN ENUMERATION THAT NUMBERS ALL THESE POINTS THOUGH THIS SECTION?]

Analyses of learning and practice are free a priori to use any index of performance: eg, errors or performance time, which decrease with practice; or amount or quality attained, which increase with practice. However, we will 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.

As we will illustrate later, the log-log law also appears to hold for learning curves defined on other performance criterion. Though significant for understanding the cause of the power law, we will only note the existence of these other curves.

Snoddy used an indicator, 1/(Time+Errors), and we have plotted log(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 trade off. He fixed the error rate to be one error per 10 seconds [CHECK]; and had the subject work faster or slower in order to hold the error rate constant. 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. Thus, Snoddy's curve actually reinforces the point that the law holds in situations where learning shows up purely as changes in performance time.

Several other things can be noted in Figure SNODDY, which will 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 relative large number of points available.

- The local deviations of the points from the line are not random. In particular, there appears to be some autocorrelation.

- The curves are for an individual, not for grouped data. This is not a condition of the law, but shows that it holds for individual data.

- 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 DATASUMMARY we tabulate several critical features of the Snoddy's data. The following equation describes the power law in log-log space:

$$\log(T) = \log(B) - \alpha\log(N) \qquad (1)$$

B is the performance time on the first trial (N = 1) and $\alpha$ is the slope of the line, ie, the learning rate. A positive value of $\alpha$, eg, .23 for the curve of Figure SNODDY, indicates a decreasing curve, since we have located the minus sign in the equation itself.

Figure SMITH&SMITH shows another series of studies on star-tracing. As with all studies in this section, log(T), the time to perform the task, is plotted agains log(N), the trial number. There are three curves, each involving a different condition of feedback. We see here another feature that we will come across several times:

- Given a family of related tasks, they often seem to *converge*, the tasks that are initially more difficult (ie, longer performance time), showing the faster learning.

The truly peculiar feature of this convergence is that if the law continued to hold, the curves would cross and the initially easier task would become the more difficult. We have extended the lines in the figure to show the point of convergence. It occurs at trial 25 [CHECK], and is approximately coincient for all three curves

[CHECK]. It is a pity that Smith & Smith did not continue the experiment to see whether or not the crossing would obtain. Even more peculiarly, it turns out that we have no example that shows the actual crossing, though we have many that show convergence and threaten to cross at some future trial. Often the crossing trial number is very large, as befits a logarithmic scale, but sometimes, as here, only a few more trials seem necessary.

A final example from a task that appears to involve intimate motor-perceptual coordination is shown in Figure CROSSMAN. 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 50 million cigars. Also, there is a known lower bound for the performance time, namely the cycle time of the machine at 6.5 sec. The curve eventually deviates from the log-log line, flattening out in submission to physical necessity. Still, practice followed the law for almost 4 million trials (and 3 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, ie, different individuals with different amounts of experience make up each point.

## 2.2. Perception

Figure KOLERS shows some data from Kolers's well known studies on reading graphically transformed text (Kolers, 1975). Here, the transformation is inversion around the horizontal axis. The task of the subject is simply 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 Koler's control data on this). In any event, as the figure shows, learning is log-log linear. The figure shows eight subjects. The scatter for the subject whose curve runs in middle is shown (OL); only the regression lines are shown for the other. This family also shows the convergence of the several curves to a point at about [XX] pages. Here, however, the family is over *subjects*, not over *task variations*, and convergence (or crossing) implies different things in the two cases.

As do some of the other examples, the experiments of Koler's exhibit many fascinating properties. In Kolers's case, there are interesting transfer effects among the family of tasks defined by different geometric transformations, and in long term learning of individual pages of text a year later. These are all important phenomena for the specific tasks. Some of them may even be important in understanding why the log-log linearity holds. But we concentrate here just on presenting the central regularity. It will be appropriate to return to these co-phenomena after the puzzle of explaining the law has taken shape.

Figure RABBITTCV shows some data from a perceptual search task (Rabbitt, Cumming & Vyas, 1979). The subject looked at a display containing a horizontal string of 3, 6 or 9 letters, to identify which target letter occurred there. The target letter was taken from a set of 2, 4 or 6 letters, and one and only one target letter

appeared in the string (at a random position). Thus, there were nine experimental conditions (3 display sizes time 3 target-set sizes). The subject worked on all these tasks together for 30 days, each day doing two rounds of a block of each of the conditions, each block consisting of 200 trials.

The plotted data is aggregated by five day periods, raising one more general methodological feature of the data sets we have to offer:

- The raw data is often arithmetically averaged over blocks of trials (occasionally over sets of subjects); whereas treatment in log-log space ideally requires geometric averaging (ie, arithmetic averaging of the logs).

In the case at hand, as in most of the others, the raw data is not available for a proper treatment of the data. It is unclear how bad the distortion is. The fits in Figure RABBITTCV appear to have some systematic departure from linearity, though it can hardly be due to statistical artifact.

The task of Rabbitt, Cummings & Vyas is essentially the same as the task investigated earlier by Neisser of finding any of multiple targets in pages of letters. Its result confirms the earlier results: with practice identification time becomes essentially independent of the size of the target set. As Figure NEISSER shows, the original Neisser data also follows the log-log law. The five curves represent two scanning for one target, two canning for five targets and one scanning for ten targets. Each curve is the average of six subjects. These points are simply replotted from the figure in the paper (Neisser, Novick & Lazar, 1963). We exhibit them, in addition to those of Rabbitt, Cummings & Vyas, to point out that much learning data in the literature fits the log-log law, even though it has not been plotted that way.

### 2.3. Motor Behavior

We have not come across any tasks with a strong motor component and a genuinely minimal perceptual component that give data relevant to the log-log law. So we offer in this section only the evidence that simple pointing shows log-log linear practice. Figure CARDEB 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). The four curves correspond to four different pointing devices: 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 speeded keys. [CHECK] Some of these devices are well described by Fitts's Law (Fitts, 1954); some have a different structure. In all cases 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. [TEXT NEEDS IMPROVING]. Also, these learning rates ($\alpha < .1$) are the smallest rates we shall encounter.

[WHAT ABOUT RUNNING TIMES AND OTHER MUCH MORE PHYSICAL OPERATIONS?]

### 2.4. Elementary Decisions

Figure SEIBEL is from a task designed by Seibel (1962) 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 stimulus 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 achieve 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; approximately the same behavior was shown by each of three subjects.

Figure BLACK shows the learning of two subjects doing elementary mental arithmetic (Blackburn, 1936). The subjects were given two digits and stated the sum. [CHECK DETAILS]. [MORE?]

### 2.5. Memory

Figure ANDERSON is from some unpublished work of John Anderson (Note 1). 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.

### 2.6. Complex Routines

Figure MORAN is from some unpublished work by Tom Moran (Note 3), done in connection with a general attack on understanding user-computer interaction. A specific, complex on-line editing task of completely rearranging a given sentence of four clauses is being performed repeatedly. The task is absolutely identical each time, ie, the same sentence. Thus we are seeing a subject simply follow an internally known, complex plan. The top curve is the total time to perform the task; The lower 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 keystrokes, rather than decreasing think time, error times, etc. Both curves show log-log linear practice effects. This figure shows one more feature that is apparent is many curves:

- Many effects can be seen superimposed on the practice curve, such as the cyclic effects of warm-up, fatigue and termination, or single-shot effects such as vacation periods. [TRUE -- CAN THEY BE SEEN IN OUR PLOT?]

[MORAN NOW (9 MAY 80) HAS SOME CURVES THAT CROSS. THIS IS OVER SUBJECTS, NOT OVER TASKS]

Figure NEVESANDERSON shows a more complex cognitive task, but one that still can be considered as evolving toward a complex routine. This is from unpublished work by David Neves and John Anderson (Note 4). 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. In fact, as we show in Figure NEVESANDERSON2, not only does the total time follow the law, but so does the number of accessing operations used. These operations are recorded automatically, because each view of a proof step, axiom or inference rule requires an explicit request for display by the subject.

### 2.7. Problem Solving

Figure STAIR 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 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*, which are initially empty. The aim is to build four *stacks*, Ace to King, 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 care is the next lower in the same suit (eg, the six of spades appended to the seven of spades).

The game can be seen to be one of perfect information. 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 of 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 is the proportion of games won. There is of course only

one series of trials, since all games, won or lost, contribute to practice. 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 trade off 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.

Since we had access to the trial by trial data, the averaging in this figure is done in log-log space (ie, geometric averages). [WAS IT?]

## 2.8. Other Tasks

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 STEVENSSAVIN is reproduced from Stevens and Savin (1962). It plots eight tasks with various response measures in log-log space. The criterion are all oriented to increase with practice. The plot is actually of the *cumulated* responses, ie, the integral of the usual curve. This is just the same as the usual power law, since 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).

$$\int_1^N Bx^{-\alpha}\,dx = B(1-\alpha)^{-1}(N^{1-\alpha} - 1) \qquad (2)$$

Some of these curves are time curves (actually, amount accomplished per unit time, to make them positive curves); but several are not, eg, #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. [SECOND THOUGHTS: MIGHT BE BETTER TO JUST GET ONE NONSENSE LEARNING AND DO IT, OR SOME VERBAL LEARNING ONES FROM MazuH78]

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, eg, "... each time the quantity of [air]planes is doubled, the cumulative average man-hours per plane will be [reduced by] 80%" (Rigon, 1944). Figure HIRSCH shows an empirical curve from machine tool

manufacture (Hirsch, 1952). Notice that the index of performance is not time, but cost.

## 2.9. Summary

We have shown some 15 [WHATEVER] diverse examples of the log-log linear law of practice. From the Table DATASUMMARY we can make one more particular point:

- The learning rates, $\alpha$, are all less than one. Indeed they cluster around a much lower value of .2 to .3, with only a very few above .5 or below .1.

In summary, 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 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 claim here that all learning is log-log linear. Nor do we claim that practice always leads to learning. The experimental literature is full of examples of measured aspects of behavior that do not improve. For example, one theme in the literature on automatization is finding conditions under which no learning occurs. The *varied mapping* conditions of Schneider & Shiffrin (1977) provide a well known instance [JA QUESTIONS THIS], but there are others. Since in this section, we have not yet attempted to structure the search for an explanation of the log-log linear learning law, it does not seem useful to exhibit negative instances. The first objective is to be impressed with the phenomena, not with its absence.

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 will just create trouble for the more cognitive instances, or vice versa, etc.

One is immediately reminded of other examples of ubiquitous regularities and their explanation. The *normal distribution*, which arises out of 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 entirely 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.

# 3. BASICS ABOUT POWER LAWS

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

## 3.1. Differential Forms and Rates of Change

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

$$T = BN^{-\alpha} \tag{3}$$

$$\log(T) = \log(B) - \alpha\log(N) \tag{4}$$

It is instructive to see this in terms of the local rate of learning, $dT/dN$.[3]

$$dT/dN = -\alpha BN^{-\alpha - 1} \tag{5}$$

$$= -\alpha T/N = -(\alpha/N)T \tag{6}$$

$$= -\alpha B^{-1/\alpha} T^{1 + 1/\alpha} \tag{7}$$

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 is proportional to T, ie, to the sum of the changes to each part of T, which on the simplest independence assumption, are the same on average:

$$dT/dN = -\alpha T \tag{8}$$

$$T = Be^{-\alpha N} \tag{9}$$

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

$$dT/dN = -\alpha' T, \quad \text{where } \alpha' = \alpha/N \tag{10}$$

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 BASICCURVESREGULAR shows these two curves in linear coordinates, with the same value of the initial value (B) taken to be 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 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 get a different law:

$$dT/dN = (\alpha/N^{\beta})T, \quad \text{where } \beta \neq 1 \tag{11}$$

---

[3] 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. [ADD IF NEEDED: We note explicitly the few places where discreteness does make a difference.]

$$T = Be^{-\alpha N^{1-\beta}} \tag{12}$$

A representative curve for $\beta$ less than 1 is shown in Figure BASICCURVESREGULAR, which produces asymptoting between the exponential and the power law.

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 $\delta$ in the first N trials, it will take another $N^2 - N$ trials to decrease by a factor of $\delta$ 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 $\delta$ in the first N trials, it will take another N trials to decrease by a factor of $\delta$ again.

### 3.2. Asymptotes and Prior Experience

As given in Equation 3, the law assumes (1) the asymptote of the learning is 0, ie, 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:

$$T = A + B(N + E)^{-\alpha} \tag{13}$$

A ($\geq$0) is the *asymptote* of learning as N increases indefinitely. E ($\geq$0) is the number of trials of learning that occurred prior to the first trial as measured, ie, prior *experience*; it thus identifies the true *starting point* of learning. (Neither $A < 0$ or $E < 0$ make any immediate sense, given these interpretations; $A = 0$, $E = 0$ reproduces the basic form of Equation 3.)

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:

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

$$\log(T) = \log(B) - \log(1 - A/T) - \alpha\log(N) - \alpha\log(1 + E/N) \tag{15}$$

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 A distort it for high N.

Figure GENERALPOWERREGULAR shows a power law with a starting point of -E and a time asymptote of A. [THE FIG HAS THE ASYMPTOTES DRAWN IN]. Figure GENERALPOWERLOGLOG shows the same curve in log-log space. Characteristically, the starting point pulls the initial segment of the curve down towards the horizontal and the finite asymptote pulls the high N tail of the curve up towards 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. [HAS SLANT LINE OF SLOPE $\alpha$ DRAWN IN.]

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

$$d(\log(T))/d(\log(N)) = -\alpha\,(1 - A/T) / (1 + E/N) \qquad (16)$$

It can be seen that the slope is everywhere smaller than $\alpha$, and becomes increasingly so as either A or E increases. A reasonable estimate of the apparent slope as viewed on the graph, $\alpha^*$, is at the inflection point. It is easy to obtain by setting the derivative of Equation 16 to zero:

$$d/dN[d(\log(T)/d(\log(N)] = -(E/N - \alpha A/T)\,(1 - A/T)\,(1 + E/N)^{-2} = 0 \qquad (17)$$

$$\alpha^* = (\alpha N^* - E) / (N^* + E) \qquad (18)$$

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

$$N^* = [BE/\alpha A]^{1/(1+\alpha)} \quad \text{where } E/N^* << \alpha < 1 \qquad (19)$$

The structure of Figure GENERALPOWERLOGLOG suggests that many of the deviations in the empirical curves could be due simply to starting point or asymptote effects. Since the effect of these two phenomenon is to bend towards 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 Figure SNODDY 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, since raising the asymptote will pull the right hand part of the curve down, and make its slope steeper. The Seibel data in Figure SEIBEL provides an example where the deviation from linearity occurs at the right. This could also be due to a positive asymptote.

### 3.3. 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 non-linear 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:

$$t(N) = T_0 + \sum_{i=1}^{N} T_i \quad = T_0 + \sum_{i=1}^{N} Bi^{-\alpha} = T_0 + B\sum_{i=1}^{N} i^{-\alpha} \tag{20}$$

$T_0$ 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 3) to be transformed to yield T(t). Instead, we proceed indirectly by means of the differential forms. From Equation 20 we obtain:

$$dt/dN = T \tag{21}$$

(Think of the corresponding integral formulation, $d/dz(\int_a^z f(x)dx = f(z))$.

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

$$dT/dt = (dT/dN) / (dt/dN) = (-\alpha T/N) / (T) = -\alpha/N \tag{22}$$

But from the basic equation (3):

$$N = (T/B)^{-1/\alpha} \tag{23}$$

Thus, we get the trials power law re-expressed in terms of time:

$$dT/dt = -\alpha B^{-1/\alpha} T^{1/\alpha} \tag{24}$$

As we saw in Equation 6, this is still the differential form of a power law:

$$T^{-(1-\alpha)/\alpha} = (1 - \alpha)B^{-1/\alpha} t + C \tag{25}$$

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

$$T = B't^{-\alpha/(1-\alpha)} \tag{26}$$

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.

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). Also, an exponential learning function in time transforms to a hyperbolic function in trials, ie, the power function with $\alpha = 1$.

[ADDITIONAL FACTS ABOUT THE POWER TRANSFORM. IF ONE NORMALIZES TIME TO BE IN UNITS OF B, ONE GETS:

$$(T/B)^{-(1-\alpha)/\alpha} = (1 - \alpha)(t/B) + C' \tag{27}$$

IF ONE NOW USES THE OBVIOUS SPECIAL CASE IN WHICH $T_0 = 0$ IN EQUATION 20, SO THAT AT $N = 1$, $T_1 = B$ AND THEREFORE $t = T_1 = B$, WE GET:

$$(T/A) = [(1 - \alpha)(t/A) + \alpha]^{-\alpha/(1-\alpha)} \tag{28}$$

I PROBABLY SHOULD INCLUDE HERE THE OTHER TWO DERIVATIONS IF NEEDED, JUST TO KEEP THEM IN A SAFE PLACE.]

### 3.4. Invariances and Compositions

Finally two general aspects of the power law are often noted as detracting from its suitability to be a law of learning.

First is its history dependence. That is, the power law is not invariant under translation. We have already seen the equations that express this: Adopting a starting point other than N = 1 (ie, setting E > 0 in Equation 13), does not reproduce a simple power law. Since, in general it seems highly unlikely that all experience prior to the official first trial of an experiment is irrelevant, it seems equally unlikely that the plot in log-log space should be a straight line, indicating that the experimental N = 1 was indeed the first trial of learning. (As noted in a prior footnote, the exponential does have the desired invariance.) Either this lack of invariance can simply be accepted as the way of the world, though perhaps surprising; or the difficulty can be buried in approximation, ie, wide ranges of starting points will all appear sufficiently like straight lines given the actual learning data.

Second, the power law does not compose well. Fundamental to the information processing view of psychology is that macroscopic behavior results from of the combined effect of component processes, eg, programs stages. The most common composition is serial, which leads to additive processes

$$T = \sum_{i=1}^{m} T_i \quad = \sum_{i=1}^{m} A_i + B_i(N + E_i)^{-\alpha_i} \tag{29}$$

Only if all the $E_i$ are the same and all the $\alpha_i$ are the same is the result a power law. The same holds true if parallel combination is considered, which leads to the Max as the composition function. This difficulty can possibly be buried in approximation.

## 4. FITTING THE DATA TO A FAMILY OF CURVES

Given empirical curves, such as occur in abundance in Section 2, it is important to understand how well they are described by curves of a given family (eg, 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 that power functions. Curve fitting without benefit of a model is notoriously a black art, so definitive answers cannot be expected, just some indication of the lay of the land.

The basic issue can be introduced from Seibel's own treatment of his data (Figure SEIBEL), which appears to be an extremely good fit to the log-log law over an extensive range (50,000 trials). Seibel (Seibel, 1963) fit his points to three curves by least squares: (1) a power law with asymptote only (ie, E fixed at 0); (2) an exponential with asymptote; and (3) a general power law with both asymptote and starting point.[4] 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, since attempts to filter out the high frequency noise prior to data fitting (say) simply add 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 sigmoid curve (ie, 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 (ie, 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

---

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

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 tactic we will attempt has four principles:

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 *spaces* of the given family. There may be more than one space.

2. Accept a curve for a family, if it plots as a straight line in the space of that family. Reject it, if it has significant shape distortion.

3. Expect a curve for a family to be rejected for alternative families in the space of those families.

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.

These criteria contain elements both of acceptance and rejection, and provide a mixture of absolute judgments about whether data belongs to a given family and relative judgments about the discrimination between families.

We will 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 will attempt to show that the both the general power and the hyperbolic family provide adequate representation of the empirical curves. Furthermore, there is no way to decide between them on grounds of theory-free data fitting. There is no space to provide a detailed examination of these techniques or of their results over the whole data set. But we do need to illustrate them enough to support the conclusions. For both results have consequences for the theoretical enterprise.

### 4.1. The Exponential Family

Figure EXPLOGLOG shows what exponential curves (with asymptote) look like in log-log space. The tail decreases too rapidly, so that at some point it falls away from the straight line. As the curve approaches the line of the positive asymptote, the downward curve must eventually become horizontal again. This turn happens quite rapidly.

There is not the least trace of a downward trend in the Seibel data. So why did Seibel get a good fit with an exponential curve? The answer is to be found in the *Achilles heel* of curve fitting -- what is the proper measure of error? The use of different coordinate systems, T-N vs log(T)-log(N) indicates that, even if we agree on a least-square criterion, we have our choice of which space to use it in. The difference is highly consequential. In T-N space, the exponential vs power differ hardly at all in accumulated error in the tail -- they both soon arrive so close to the asymptote that the squared error is negligible, compared to errors at low

values of N. In log-log space, on the contrary, exponentially increasing weight is given to the tails. Seibel did all of this fitting in T-N space and could see little difference. In log-log space there is little evidence for an exponential. Another error measurement that has similar properties to the least-square measure in log-log space is the relative least-squares criterion. Each residual is divided by the value of T before it is squared. Relative to the standard measure, this has the effect of increasing the weight of the asymptotic portion. We have used this measure extensively whenever data is to be fit in a non-logarithmic space [STILL NEED TO UNDERSTAND WHETHER THIS ANALYSIS IS RIGHT, AND IF NOT WHAT DOES EXPLAIN IT. SHOULD REALLY LOOK AT WHAT THE BEST FITTING EXPONENTIAL CURVE FOR SEIBEL LOOKS LIKE IN LOG-LOG SPACE, AND SEE HOW NON-LINEAR THAT IS]

We need to plot Siebel's data points in a space that makes exponential curves into straight lines. The space normally used comes from equation 9 for the exponential:

$$\log(T) = \log(B) - \alpha N \tag{30}$$

This plots as a straight line in log(T)-N space. Figure SEIBELEXP shows this plot for Seibel's data. The curve does not follow a straight line at all, but moves away from it toward the right. This is exactly what one would expect from a power law, ie, the tail is much flatter. Of course, this is also what one would expect if there was a positive asymptote. The existence of an asymptote destroys the straight line in log(T)-N space. A search could be made for the asymptote which minimizes the deviation in log(T)-N space. The results of this search can be seen in Figure SEIBELEXPA. [PLACE THE RESULTS OF THIS ANALYSIS HERE. I'M NOT SURE WE CAN GET AROUND THE PROBLEM OF A HAVING TO BE LESS THAN EVERY VALUE OF T, BUT THIS ANALYSIS SHOULD BE DONE. INCLUDE A COMPARISON OF THE DEVIATIONS WITH A POWER LAW PLOTTED IN THIS SPACE].

There is an alternative space in which the existence of an asymptote does not destroy the linearity. This space comes from the original equation:

$$T = A + BX \qquad \text{where } X = e^{-\alpha N} \tag{31}$$

We will refer to this as the T-X space for the exponential. Figure SEIBELET-X shows Seibel's data in this space. In order to be able to plot in the T-X space, a value for $\alpha$ must be assigned ahead of time. Once this has been done, the transformation from N to X can be made, and the plot can be generated. The $\alpha$ used in figure SEIBELET-X is 0.000106, as this value was determined to yield the best fit (see Table DATASUMMARY). There are shape distortions in this curve which imply that Seibel's data are not fit well by an exponential. By comparing this graph with one for a power law plotted in exponential T-X space (Figure GENERALPOWERET-X), it becomes apparent that these deviations are what would be expected of a power law.

For both of these analyses we had to search for one of the parameters. The fundamental problem is that

there are three parameters, A, B, $\alpha$, and no two dimensional space exists that permits a plot of the curves as straight lines with all 3 parameters dropping out of the analysis. Thus, we have to take our choice of two and then search for the third (or third & fourth for general power). Thus for exponential we could take log(T)-N and search for A, or T-X and search for $\alpha$. It should not matter which space is used, since significant deviation in shape, in either space, would signify a poor fit by an exponential. Though this correlates with error measurements, the standard error measurements are not good in distinguishing shape distortion from random fluctuation.

Whichever space is used [ASSUMING THAT THE SEARCH IN LOG(T)-N SPACE COMES OUT RIGHT], it can be seen that Seibel's data are not well fit by an exponential. We have examined all of the curves in Section 2 in this fashion. The results are indicated in Figure(DataSummary) in the column labeled *Exponential.* [USE SOME CODE TO SHOW HOW STRONGLY IT INDICATES EXPONENTIAL OR POWER IN LOG(T)-N SPACE] We conclude that indeed it is reasonable generally to thing of this data as being a power law to a first approximation.

### 4.2. The General Power Family

As with the exponential family, there are several spaces in which the appropriateness of power laws can be judged. There is the log(T)-Log(N) space. Figure SEIBEL shows Seibel's data plotted in this space. The curve is sufficiently linear that it is not necessary to search for the best A and E in order to determine that a power law is a reasonable model of the data. This impression is confirmed by looking at the data in the power law T-X space (Figure SEIBELPT-X). This space is defined by:

$$T = A + BX \qquad \text{where } X = (N + E)^{-\alpha} \tag{32}$$

A fit in this space requires giving values to both E and $\alpha$. The optimal values for Seibel's data (determined by a two dimensional search) can be found in Table DATASUMMARY.

[SHOULD PROBABLY HAVE A COMMENT IN HERE ABOUT THE SEARCH, AND THE CHARACTERISTICS OF THE E-$\alpha$-SSD SPACE. POINTS TO FIGURES SEIBELALPHA AND SEIBELE]

### 4.3. The Hyperbolic Family

The hyperbolic family is:

$$T = A + B/(N + E) = A(N + B')/(N + E) \tag{33}$$

The equation can be taken as a special case of the general power law (Equation 13), with $\alpha = 1$, or as a family in its own right, with three parameters, A, B and E, which have the same interpretation as in the general power law.

The hyperbolic has been put forth as the universal form of the learning curve (Mazur & Hastie, 1978, Lippert, 1976), in much the same spirit as we, following the human skill literature, have put forth the power law.

In log(T)-log(N) space a hyperbolic should be linear with a slope of -1, but the apparent slope of Seibel's data is approximately -0.3. This discrepancy does not rule out a hyperbolic model though. As was seen in equation 18, the perceived slope in log(T)-log(N) space $-\alpha^*$ is shallower than the actual slope $-\alpha$. If A and E are not 0, then the curve could be a hyperbolic, and still achieve a much shallower slope in log(T)-log(N) space. Doing a straightforward two dimensional search over A and E could not accomplish this though, because in log(T)-log(N) space $\alpha$ is a parameter determined by regression. It cannot be forced to be 1 like it must be for a hyperbolic. The resolution to this problem is to look in the hyperbolic T-X space, defined by:

$$T = A + BX \qquad \text{where } X = 1/(N + E) \tag{34}$$

Figure SEIBELHT-X reveals that a hyperbolic is a reasonable model of Seibel's data.

### 4.4. Summary and a Practical Note

We believe that we have established the reasonableness of excluding the possibility that practice learning is exponential, the reasonableness of describing the data by power laws, and the unreasonableness of excluding either the hyperbolic family as sufficient or the general power family as necessary [WE NEED TO BACK THIS UP MORE WITH SOME EVIDENCE AS TO WHY A POWER LAW MIGHT BE NECESSARY, SUCH AS CURVES IN WHICH THE OPTIMUM $\alpha$ IS FAR FROM 1]. It would be nice to more precise about the latter, but the data we have considered to not allow it.

A side note is appropriate about the practical side of the matter. There remains a great virtue about plotting practice data in log-log space, ie, plotting log(T) versus log(N) untransformed. This provides a technique that permits useful analysis of the data without requiring the estimate of parameters. For instance, though a hyperbolic may seem like the useful family, with is no reason to go into log-log space, there is no way to obtain the parameters without an estimation procedure of some kind. It will not, of course, plot as a straight line in log-log space with $\alpha = 1$. But with the principles developed here, it is possible to interpret the plot in log-log space directly to yield the parameters and to interpret the nature of the curve.

As we have developed, $\alpha^*$, the apparent $\alpha$ will be flatter that the true $\alpha$ depending on the values of E and A. In fact, the following relations hold:

$$E = N^*(\alpha - \alpha^*) / (1 + \alpha^*) \tag{35}$$

$$A = T^*(\alpha - \alpha^*) / \alpha(1 + \alpha^*) \tag{36}$$

If it is assumed that the family is the hyperbolic, then $\alpha = 1$ by stipulation. Then we get:

$$E_{hyperbola} = N^*(1 - \alpha^*) / (1 + \alpha^*) \tag{37}$$

$$A_{hyperbola} = T^*(1 - \alpha^*) / (1 + \alpha^*) \qquad (38)$$

Thus, these provide handy tools for making estimates directly from the log-log plot to the values of E and A. [NOT QUITE TRUE -- STILL TO BE FINISHED. ALSO JA CLAIMS THIS DOESN'T MAKE SENSE.]

# 5. 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, and possibly even more widely.

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

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 Section 2 that such a widespread phenomenon can only result from some equally widespread structural feature. Possibly, a mixture of learning mechanisms yields something approaching a power law. The second 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? The third 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.

As noted initially, we will concentrate on a single explanation. 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.

## 5.1. General Mixtures

The following qualitative argument has a certain appeal.

> *The Mixtures Argument*: Performance depends on a collection of mechanisms in some monotone way -- ie, 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, too boot), so the components they effect 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.

A natural class are the exponential functions. They form a rich enough class (a three parameter family of $\alpha$, A and B). They also are as good a candidate as any for primitive learning functions.

Consider first additive systems, ie, where each component adds it contribution to the total performance. This means that T is a weighted sum of exponentials.

$$T = \sum_i W_i e^{-\mu_i N} \tag{39}$$

Figure EXPONENTSUM shows a plot in log-log space of a <four> term sum with weights (the W's) and rates (the $\mu$s) selected at random. [OR SOMETHING]. One gets a reasonable approximation to a straight line, though it is a little wavy. [OR WHATEVER]

However, three arguments should make us somewhat wary that such mixtures can work in general. [NOT SURE WHAT THE RIGHT ORDER IS FOR THE THREE]

### 5.1.1. Piecewise independence

Consider two separated parts of the learning curve, say the part labeled #1 and the part labeled #2 in Figure EXPONENTSUM. Grant for the moment that sums of exponentials lead to straight lines in log-log space. Then why should the slope of the straight line over part #1 be the same as the slope over part #2? The separation guarantees that the straight line gets its significant contributions from different members of the total population of exponentials. There is no reason they should lead to the same composite learning rate.

Conceivably, some other underlying structural feature conspires to fix all composite rates at a common value. If true, the source of such a regularity is an essential ingredient to the law and the general law of mixtures does not suffice by itself.

Such a regularity would imply that all empirical learning rates should be the same in log-log space. A glance at Table DATASUMMARY shows, interestingly enough, both both substantial variation and strong limitation of range. Mostly, $\alpha$ is between .2 and .3, and in no case is it greater than .5. [OR WHATEVER]

### 5.1.2. The pure tail

Consider a finite collection of exponential learning mechanisms, ie, a finite sum in Equation 39. The rapid learners will certainly drop out, in agreement with the argument. However, what is left is the final exponential. Thus, the tail of the total learning curve comes to be identical with the tail of the slowest component. Which is to say, it becomes exponential, rather than a power law. Thus, a mixture works only in some initial range. Correspondingly, this means that by some happenstance our empirical curves would have

to have been limited to this initial range, even those like Seibel (Figure SEIBEL) and Crossman (Figure CROSSMAN) that go many trials. There are surely deviations towards the end of these curves; but unfortunately, they go in the opposite direction from an exponential tail (which would curve downward).

### 5.1.3. The representational power of exponentials

The most important argument is simply that sums of exponentials 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:

$$T(N) = \int_0^\infty W(\mu)e^{-\mu N}\, d\mu \tag{40}$$

On the one hand, this simply expresses the continuous analog of a sum of exponentials: the exponential for every $\mu$ is represented, each with its own weight, $W(\mu)$. On the other hand, this will instantly 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 from any function T(t) there is a function $W(\mu)$ that produces it.[5] Thus, by choosing appropriate weights, any total learning function whatsoever can be obtained.

We can of course choose weights to make T a power law, as in Equation 3, with $\alpha$ and B. Consulting any standard table of Laplace Transforms shows:

$$W = (B/\Gamma(\alpha))\mu^{-(1-\alpha)} \tag{41}$$

That is:

$$T = BN^{-\alpha} = \int_0^\infty (B/\Gamma(\alpha))\mu^{-(1-\alpha)}e^{-\mu N}\, d\mu \tag{42}$$

The component exponentials correspond to learning at all rates, indefinitely fast (large $\mu$) to indefinitely slow (small $\mu$). Since $(1 - \alpha) \geq 0$, the weight W becomes very small for fast learning and very large for slow learning. Since there is no particular reason to expect such a distribution of weights on the components, it would seem implausible that power laws would be produced by such mixtures.

However, we can turn the argument around and get a positive result. One distribution of weights that is natural is the rectangular, ie, 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 can be seen from Equations 41 and 42, this corresponds to $(1 - \alpha) = 0$, which yields $\alpha = 1$. The resulting law is the hyperbolic.

If we examine this result with respect to our arguments against mixtures we see that it does indeed have an additional feature that assures that the slope stays the same throughout the entire range, namely, the

---

[5] T must be mathematically well behaved in certain ways to be so represented, but these are of no consequence in the present context.

rectangularity of the weighting. [THIS DOESN'T SEEM COMPLETELY RIGHT] That this is an assumption of disorder, not order, is what lets us make it without any additional substantive commitment. Furthermore, exponentials of *all* rates are involved, so there is no pure tail.

It is beyond the bounds of this paper 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 good candidate descriptive curve, this adds up to a significant observation (it can hardly be distinguished as a "result").

### 5.1.4. Mixing other ways.

Simple additive combination is not the only way to put learning mechanisms together. For instance, Clayton Lewis (Note 2) 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 [DID HE IN THE PAPER? -- AT LEAST IN PERSONAL COMMUNICATION] at a position roughly in agreement with the remark in the subsection on piecewise independence that another source of constraint or uniformity is needed.

[WHAT ABOUT MIXTURES OF POWERS -- THEY ARE JUST LAPLACE TRANSFORMS TO ON LOG(N) INSTEAD OF N. ]

### 5.1.5. Summary

If we review the three difficulties discussed above -- piecewise independence, pure tails, and representational power -- we see that all of them are likely to dog any form of mixture, without some additional strong constraint, which constraint would actually carry the determination of the integrated form, eg, to a power or some other function. The one exception we have come on is the generation of the hyperbolic as a mixture of exponentials under the assumption of a uniform initial amplitude of exponential component processes.

[SHOULD WE DO MORE ON LEWIS'S SCHEME?]

## 5.2. Exhaustion of Exponential Learning

There are several ideas for exhaustion, depending on what it is that becomes depleted as learning proceeds.

### 5.2.1. Method exhaustion

Suppose there were a pool of possible methods for performing a task. Learning would consist of selecting a new method in the pool and using it instead of the existing method. Without some assumptions, nothing can be said about how the learning would proceed. Even to get the next method always to be an improvement,

requires some assumptions on the subject's ability to analyze the potential effects of methods. However, by making various simple assumptions we can begin to see what sorts of learning laws method selection can easily yield.

The reasonable basic assumption is that the change in performance of a method is proportional to the time to perform the method. Given two methods, one taking 1 sec to perform, the other 10 secs, we expect the learning of the 1 sec method to be smaller than of the 10 sec method, other things being equal. This basic assumption leads of course to the exponential (as in Equations 8 and 9).

To be a power law, there needs to be something to slow down this process. One possibility, among others, is that the good methods get used early in learning, leaving only poorer methods. This we can refer to a *method exhaustion.* Certainly, the form of the differential equation for the power law (Equation 6), can be seen this way -- the rate of learning, which is the average effectiveness of the new method for each part of the performance time, continually diminishes. The problem is to find mechanisms that make it diminish by 1/N.

[I THINK ROSENBLOOM HAS A MODEL THAT YIELDS HYPERBOLIC. IT SHOULD GO HERE.]

### 5.2.2. Search exhaustion

The quality of methods is not the only thing that can be exhausted. Let the methods have a range of improvements, but with no ability on the part of the subject to select them differentially. However, let it be necessary to search for a method. Then it seems reasonable that there could be *search exhaustion*, in that method improvements are immediately at hand initially, but take increasingly long to find. Therefore, in some way the rate of improvement, which is the product of the average size of improvement time the chance that an improvement will be found, will decrease steadily.

[DO WE HAVE AN ACTUAL MODEL?]

### 5.2.3. Parallelism: Exhaustion of contribution

One long standing view of learning is that of an initial form that is deliberate, conscious and resource limited, with a gradual transformation to a form that is automatic, unconscious and resource independent. One image of this in mechanism is that learning consists of a transformation from a serial to a parallel processing organization.

Under suitable assumptions, such a transformation might yield an appropriate law of learning. The underlying idea is that, as the performance process becomes more parallel, the learning effort must still be distributed over the entire parallel ensemble, but only a fraction of this is effective in decreasing the duration of the process.

Specifically, let there be an amount of processing (ie, operations) P that must be performed. Initially, it all

happens in series, so it takes time P. Learning creates parallel processing. To think of this in continuous terms, let the process have a given width W and length, T. Then, since the same total processing must be accomplished, at all trials

$$P = T * W \tag{43}$$

Learning at any moment (dT/dN) is basically proportional to the amount of time available (T) -- the usual exponential assumption. However, it must be spread over the entire set of parallel processes, ie, over the entire width (W). [DO WE WANT A FIGURE WITH RECTANGLES?] That is:

$$dT/dN = (\alpha/W)T \tag{44}$$

But from Equation 43:

$$dT/dN = (\alpha/P)T^2 \tag{45}$$

$$T = (\alpha/P)N^{-1} \tag{46}$$

Thus, we get a hyperbolic law instead of a power law with general $\alpha$. An essential feature of this is the conservation assumption of Equation 43 that always the same amount of total performance work (P) must be accomplished. In some sense, there is no real learning, only the reorganization to increase the parallelism.

This assumption can be changed in various ways. For instance, the learning effort might be associated with the performance process, rather than with time, thus permitting each of the parallel streams to continue to learn in parallel. The effect of this can be seen simply to leave the entire learning exponential. Perhaps, co-acting learning process could shorten the total performance (P) required. Even without positing any particular law, the effect of this can be seen to speed up learning. Since the hyperbolic equation (46) already has a slope of -1, this can only increase its effective slope (whether it does so while maintaining a power law or bends it over sharply, as in an exponential). Thus, elaborations in this direction will not move to produce a power law of the right character, namely, one that is slower than then hyperbolic.

[SHOULD WE PUT IN AN ACTUAL EXAMPLE OF COMPOUND LEARNING? WOULDN'T THINK SO. HERE IS ONE:

LET P BE DETERMINED BY ITS OWN INDEPENDENT EXPONENTIAL LAW

$$dP/dN = -\beta P \tag{47}$$

$$P = Be^{-N} \tag{48}$$

$$dT/dN = -(\alpha/P)T^2 \tag{49}$$

$$dT/dN = -(\alpha/B)e^{\beta N}T^2 \tag{50}$$

$$T = (\beta B/\alpha)e^{-\beta N} \tag{51}$$

NOTE THAT IT CONVERTS IT TO AN EXPONENTIAL THAT DOES NOT REALLY REFLECT THE PARALLELISM AT ALL. IS THIS RIGHT?]

### 5.3. Stochastic Selection

The work in stochastic modelling generated a large range of models, well beyond what we can review. Two specific examples are highly relevant and will give the flavor of the approach.

#### 5.3.1. 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 (Figures CROSSMAN and BLACKBURN are taken from his paper). He proposed a general model based on an improving process of selecting methods from a fixed population of methods. 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, $t_i$:

$$\Delta p_i = -k_1(t_i - T(N)) \tag{52}$$

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

$$T(N+1) = T(N) - k_1 \text{Variance}(\{p_1\}) \tag{53}$$

In general, the time course cannot be calculated without knowing the actual distribution of the $p_i$, for Var(n+1) will be a function of Var(n) and the third moment; the change in this moment from n to n+1 in turn depends on the next higher moment; and so on. Crossman assumed a (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 Figure GENERALPOWERLOGLOG. 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 can be seen from Equation 52, nothing prevents $p_i$ from moving outside the [0 1] interval, thereby violating the basic property of being a probability. Indeed, if the i-th 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.

A small but interesting step can be taken beyond Equation 53. For simplicity, Crossman considered an arbitrary population of methods, but methods could be taken to be related to each other, eg, composed as programs built up out of subprocesses (eg, operators). Variant methods involve few operators, though they might otherwise be quite different in composition. If the basic operators are serially uncorrelated, then the composition is like the sum of independent random variables, so that the variance is proportional to the mean. Thus, we get:

$$\text{Var}(N) = k_2 T(N) \tag{54}$$

$$T(N+1) = T(N) - k_1 k_2 T(N) = [1 - k_1 k_2]T(N) \tag{55}$$

This clearly defines T(N) as exponential learning. There is in fact some evidence that human operations show the basic compositional law of Equation 54 (Card, Moran & Newell, 1980a), besides its being a rather likely relation on a priori grounds.

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

### 5.3.2. The Accumulator and Replacement models

Among the basic stochastic learning models two classes 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.

## 5.4. Conclusion

We have considered several basic ideas that might give rise to the power law of practice. Though all these ideas have considerable plausibility, examples of power laws have not come forth easily, by formulating the simple and obvious versions of each idea. The mixtures turned out to be too powerful, ie, they could integrate too wide a range of total learning functions. Thus for them the additional constraint was missing. The exception was obtaining the hyperbolic. From the exhaustion ideas, what again emerged were hyperbolic laws, with no apparent way to enrich them from a power law with exponent of $-1$ to a general (negative) exponent (or one generally less than one). Again, from stochastic selection comes hyperbolics with ease. Crossman's model, while providing a non-linear model is does not provide a power law. Unfortunately, it is ill-formed and does not seem worth patching up.

The conclusion we wish to draw is not that power laws are actually impossible to formulate, only elusive. On the contrary, we believe that versions of each of the basic ideas we have explored can yield power laws with the appropriate formulation and a little more fiddling than we have done. We do wish to provide a background that enhances the discovery of a power law in one additional approach, which we will now consider. This is again an exhaustion idea, if viewed correctly, namely of usefulness in performance.

Likewise, we do not wish to underestimate the way the hyperbolic turns up at every turn in our investigation. That seems to us significant. However, we do want to put it to one side while we present this other model.

## 6. THE CHUNKING THEORY OF LEARNING

We take as central to our model a theme which 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, eg: (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, to make matters concrete. There are ten lights $L_1$, ..., $L_{10}$, 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*. These are first level chunks from the point of view of Seibel's task; clearly they are built up from still more primitive features, relations etc. 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 level; eg, the *height* of the last mentioned chunk, $((L_7^+\ L_8^-)\ (R_7^+\ R_8^-))$, is three levels up from the primitive chunks of $L_7$, +, $L_8$, etc. 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 task environment provides combinatorial possibilities for chunks as the height of chunking 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 multi-level 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:

$$\begin{array}{c}((L_x +)(R_y +)) \\ L_1 + \quad R_1 +\end{array}$$

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 ($L_1$ +), 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 height 6, to cover all the cases.

$$((((L_1^+ R_1^+)(L_2^- R_2^-))(L_3^+ R_3^+))\ ((L_4^+ R_4^+)(L_5^+ R_5^+))\ ((L_6^+ R_6^+)((L_7^- R_7^-)(L_8^- R_8^-))((L_9^- R_9^-)(L_{10}^+ R_{10}^+)))) \qquad (56)$$

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

*On the structure of the task environment:* An assumption about combinatorial productivity may appear unfamiliar, but it rests on an elementary feature of almost all task environments. Observe in Seibel's task (and thinking only about the lights) that there are only two patterns of one light (on and off), but four patterns for two lights, eight patterns of three lights, and so on, up to 1024 patterns of ten 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.

This multiplicity of patterns (or chunks) depends on their being an entire *ensemble* of possible concrete environments, meaning by this latter what the subject faces at a particular trial in which he performs the task. If we take a single concrete environment, presented to the subject at a particular trial, then only a small number of the possible chunks occur. Indeed, at the top-most level, the entire concrete environment at a trial can be encoded in a single chunk, as in example 56 above. But learning implies that the subject is faced with an ensemble of environments; and the trial sequence provides the sample of concrete environments actually experienced. When we refer to the *task environment*, we mean the entire ensemble.

Given a task environment, composed from a set elements which can vary with respect to attributes, locations, relations to other elements, etc., the number of possible chunks involving a pattern of these elements grows combinatorially as the size of the pattern increases. In terms of the height of chunking, each additional step of chunking multiplies the number of total chunks by some factor, as existing chunks are combined with some set of additional elements. 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. But this combinatorial structure will persist.

*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 when it is performing in a concrete environment. This is consonant with theories that have have learning occuring automatically from the chunks (involving both the stimuli and the subject's own responses) that are built in Working Memory -- when 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 that 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 (ie, had already been learned) and were being perceived in the environment.

These three assumptions, though still general, provide a basis on which a specific learning model can be built. We will start by presenting the absolutely simplest form of this model, so as to reveal its structure clearly. Various limiting conditions and the like may appear somewhat strained in this simple version. We will generalize it in the next section.

### 6.1. Simple Version of the Chunking Model of Learning

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

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

h = The height of chunking.

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

$$dT/dN = (dT/dh)\,(dh/dC)\,(dC/dN) \qquad (57)$$

The first term, dT/dh, expresses how performance time (T) changes with the height of coding. 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). Now a chunk is just a tree of subchunks, down to the primitive chunks. Let c be the average number of subchunks per chunk. Then, if chunking has proceeded to height h, each top-level chunk spans $c^h$ initial chunks. Thus, the number of top-level chunks that are required to span the performance is $P/c^h$ and we get for the performance time:

$$T = Pc^{-h} \tag{58}$$

$$dT/dh = -\log(c)Pc^{-h} = -\log(c)T \tag{59}$$

If this holds for unlimited values of h, it implies that P is indefinitely divisible and that T can be driven to zero. We just accept such simplifications in this first model.

The second term of Equation 57, dh/dC, expresses how fast the height of the chunks increases as the subject accumulates more chunks. It depends on how many chunks at each level 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, 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. Let b, for the *branchiness* of the task environment be this average factor. Then, if there are originally S primitive chunks about the environment, the total number of chunks to cover patterns of h elements in the task environment, $C_{te}$, is given by:

$$C_{te} = Sb^h \tag{60}$$

We need to relate $C_{te}(h)$, the total chunks (at level h) defined on the environment, to C(t), the total chunks that the subject has at a given time. 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. To a rough approximation, suitable for this simplest model, we can equate C and $C_{te}$. That is, if the subject has learned C chunks, these will be all the chunks provided by the environment from the elementary chunks up to the h that yields C. Hence we get:

$$C = C_{te} = Sb^h \tag{61}$$

$$dC/dh = \log(b)Sb^h = \log(b)C \tag{62}$$

$$dh/dC = 1/(\log(b)C) \tag{63}$$

The final term of Equation 57 follows directly from the assumptions on learning that the number of chunks learned per unit time is a constant, say $\lambda$ chunks:

$$dC/dt = \lambda \tag{64}$$

Therefore by Equation 21, which relates time to trials:

$$dC/dN = (dC/dt)\,(dt/dN) = \lambda T \tag{65}$$

We now have assembled all the components of Equation 57:

$$dT/dN = (-\log(c)T) \quad (1/(\log(b)C)) \quad (\lambda T) \tag{66}$$

$$= -\lambda(\log(c)/\log(b))\, C^{-1} T^2 = -(\lambda/\rho)C^{-1} T^2 \quad \text{where } \rho = \log(b)/\log(c) \tag{67}$$

It is still necessary to express C as a function of T, which can be done through the common link of h via Equations 58 and 61:

$$C = Sb^h = Se^{h\log(b)} = S(e^{-h\log(c)})^{-\rho} = S(c^{-h})^{-\rho} = S(T/P)^{-\rho} = SP^{\rho} T^{-\rho} \tag{68}$$

$$dT/dN = -(\lambda/\rho)\, S^{-1} P^{-\rho} T^{2+\rho} \tag{69}$$

We can now recognize this as the differential equation of a power law (see Equation 7), whose solution is:

$$T = [\lambda(1+\rho)/\rho]^{-1/(1+\rho)} S^{1/(1+\rho)} P^{\rho/(1+\rho)} (N + E)^{-1/(1+\rho)} \tag{70}$$

The constant of integration shows up as the variable starting point.[6] On the other hand, that the asymptote is 0 is built into the differential equation from Equation 58, which permits performance time to decrease indefinitely. Likewise, $\alpha$ and B, the free constants in the basic equations 3 and 4 for effective learning rate and initial performance time respectively, are no longer free, but are determined by the basic parameters, $\lambda$, P, S, c and b.

$$\alpha = 1/(1 + \rho) = \log(c)/(\log(b)+\log(c)) \tag{71}$$

$$B = [\lambda(1 + \rho)/\rho]^{-1/(1+\rho)} S^{1/(1+\rho)} P^{\rho/(1+\rho)} \tag{72}$$

The learning rate, $\alpha$, depends only on the relative rates of exponential growth, $\rho$, ie, the expansion factor at which the environment spawns chunks to be learned (b), versus the compression (ie, chunking) factor at which the subject can chunk them (c). It may seem odd that the rate at which the subject manufactures chunks ($\lambda$) does not enter into this learning rate, but this is typical of how linear factors ($\lambda$) cannot dent exponential processes (the hierarchical chunking controlled by b and c). Note that $\alpha$ must always be less than 1, as long as both b and c are greater than 1, ie, as long as they are positively generative. If either is less than 1, than the hierarchies close off and anamolous behavior results. [IS IT INTERESTING BEHAVIOR?]

It is simpler just to use $\alpha$ itself as the fundamental index of the complexity of the environment relative to chunking power, rather than $\rho$. Then B simplifies to be:

$$B = [\lambda/(1-\alpha)]^{-\alpha} S^{\alpha} P^{1-\alpha} = [(1-\alpha)S/\lambda P]^{\alpha} P \tag{73}$$

$$= [(1-\alpha)F]^{\alpha}P \qquad \text{where } F = S/\lambda P \tag{74}$$

This last form is perhaps the most perspicuous. The initial performance would be expected to be just P, the initial set of chunks to be processed. In Equation 74, B is indeed proportional to P. However, due to the continuity of the model, there is learning throughout the initial trial, which makes the effective P smaller. The factor of $[(1-\alpha)F]^{\alpha}$ represents this contribution. It is smaller (ie, has more effect), the larger the learning rate

---

[6] The integration is of the form $T^{\gamma}dT = \beta dN$, which yields $T^{\gamma+1}$ = constant*N +constant, so that T = (constant*N +constant)$^{1/(\gamma+1)}$ = constant(N +constant)$^{1/(\gamma+1)}$.

($\lambda$); the larger the initial performance (P) (since there is more time to learn on the initial trial); and the less complex the environment (smaller S and smaller $\alpha$).

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, ie, each chunk increases the height of chunking by one and this is enough to keep dT/dh proportional to T (Equation 59). Thus there is neither search nor method exhaustion. However, the chance that a chunk will get used becomes increasingly rare. It becomes rarer, actually, because of the height 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 (ie, low-level) chunks are learned first and specialized chunks are learned later.

We thus have one theory that yields a power law. It is, however, a highly particular and simplified version of the general assumptions stated at the beginning of the section. Two directions need to be taken: (1) explore the ramifications of the theory and connect it up with other work in cognitive psychology; and (2) determine how general a theory of this form still maintains the basic properties. We start with the second task, since a key property of the log-log law is its generality. If the theory holds only for the particular assumptions we have just made, that will be a genuine mark against it.

### 6.2. Generalizing the Model

We wish now to generalize the theory -- not the basic assumptions, but their concrete form while still implying a power law.

The specific version in the last section is too simple in a number of respects. In common with all macroscopic models, it uses averages and aggregates. For instance, the size of chunks is not necessarily uniform with level, so perhaps $c^h$ should be $c_1c_2...c_h$. Further, the actual chunks vary with many factors, so that a closer picture could be gained by a stochastic model. Similar, considerations apply to the chunks in the environment and to the learning. However, we put all considerations of this sort to one side and continue to deal with macroscopic models characterized by averaged quantities. These types of generalization are certainly important, but we are more concerned with the structural assumptions of the model. Thus, we focus instead on how the performance program might depend on chunks, how the environment might give rise to chunks, etc.

The assumptions divide into the three aspects, concerned respectively with the performance program, the structure of the task environment and the learning rate. We take up each in turn.

### 6.2.1. Performance

Performance time is surely not simply proportional to the number of top-level chunks. For one thing, the top-level chunks must be recognized and, even if they dictate action directly in terms of action-chunks of some sort, this must propogate down through the chunks in some way. On one view, the environment offers lower level features and a parsing process of some kind must aggregate upwards to produce the highest level features. Even if processing were totally parallel, this would take time proportional to h, the number of levels of chunking. Similarly, flow of control down through response chunks would seem to be at least proportional to the height of the response chunks.

Here are some complexities and some simple-minded ways they might effect total performance time (T). We are not interested in the details, but only in the form of the dependence of performance as a function of h to see what shape the generalization should take. Thus, we use the $k_i$ throughout as arbitrary positive constants within each equation.

1. The chunks that apply to the current environment must be recognized. If the recognition occurs first and if it depends on the height of coding, we might have:

$$T = k_1 h + k_2 c^{-h} \tag{75}$$

2. High-level response chunks must propogate their execution to the actual response. This is the inverse of perception (parsing) and again would depend on the height of command chunks to be decoded. These chunks might be somewhat smaller than the total height of chunking, h, but would still be of order h.

$$T = k_1 c^{-h} + k_2 h \tag{76}$$

3. The process may have inherent limits on its speed-up, with a number of chunks requiring processing in any case:

$$T = k_1 + k_2 c^{-h} \tag{77}$$

4. The task may be inherently sequential (eg, in playing solitaire, the deck must be dealt before the hand is played) and learning may occur separately on the components with unequal c and h:

$$T = \sum_i k_i c_i^{-h_i} \tag{78}$$

5. The top-level chunks of height h do not cover a sequence of steps of $c^h$, but only somewhat less, because the top several levels put together stimulus chunks with response chunks, provide sequence information within the total program, etc. For instance, the top-level chunk might look like:

   $((\text{After Step}_i)\ (\text{Stimulus}_j\ \text{Response}_k))$

   The response subchunk might be the one that actually generates coverage of steps in the task; here it is at level h – 2. Thus, in general the top few levels might not contribute to coverage.

$$T = k_1 c^{-(h-k_2)} \tag{79}$$

6. The effect of a chunk on performance may not be to aggregate correct steps, so as be able to take

big-chunk sized processing steps, but to select the correct path, eliminating errors and the processing produced by them. For instance, the processing might be expressible as:

$$T = T^{necessary} + \sum_i \pi_i T_i^{error} \tag{80}$$

Here $\pi$ is the chance of selecting the corresponding T and is affected by what chunks (knowledge) the subject has; in the simplest model knowledge sets some $\pi_i$ to 0. The chunks that hold this knowledge are of different heights. The exact distribution is dictated by the task environment, but it is not unreasonable to take it simply as uniform with h. The important thing is that the support of each effective chunk, ie, the pyramid of subchunks that lead to it, is not likely itself to be effective. Thus, effective chunks grow more as a linear function of h than an exponential one. Given this, improvement in T is linear in h, assuming that the size of the saved time (the $T^{error}$) is uncorrelated with the height of the chunks, and so can be replaced with an average time per error. Thus we get:

$$T = [k_1 + (k_2 - k_3 h)k_4]c^{-h} \tag{81}$$

[THIS EXAMPLE TO BE REVISED AND SIMPLIFIED, ESPECIALLY THE EQUATION]

The array of possibilities above suggests that performance might generally be a linear form in h and $c^{-h}$. On purely formal grounds, it might be extended to a linear form in $c^{-h}$ but polynomial in h (which however must always be positive):

$$T = P_0(h) + P_1(h)c^{-h} \quad \text{where the P(h) are positive polynomials in h} \tag{82}$$

The one constraint that is not reflected in this polynomial assumptions is 4. This expresses the general sum of power laws and the question of its approximation needs to be handled separately.

### 6.2.2. Task Structure

The task environment is to be characterized by the number of chunks, $C_{te}$, that can be defined on it for each level. Let us enumerate the chunks in a somewhat more explicit model of the environment than we considered originally. Suppose there are q objects in the environment, each of which has a attributes, each attribute of which has v values -- all these being average figures. Suppose further each object is linked to k others, in the sense that it is close enough (semantically or perceptually) to give rise to a chunk. Then the number of chunks at each level, h, is:

$$C_{te}(h+1) = kavC_{te}(h), \quad C_{te}(1) = qav \tag{83}$$

$$C_{te}(h) = (kav)^{h-1}C_{te}(1) = (q/k)(kav)^h \tag{84}$$

The chunks that are being counted here correspond to the patterns of elements in the environment. New chunks are formed if they represent new relevant structures on the environment (including responses). If the chunks per se are taken as distinguishable (a syntactic point of view, so to speak), then the number of formable chunks goes up much faster that exponential. Eg, even without any further input from the environment, given C(t) total chunks at time t, one gets all combinations of C(t) taken c at a time for C(t+1), and then all combinations of C(t+1) taken c at time again for C(t+2).

The above model is static. Consider another simple model for a dynamic situation. Suppose the task consists of k sequential steps, $S_1$, $S_2$, ..., $S_k$, each step leading to any of b possible next steps, ie, b different environments can follow any given environment. The chunks of interest are the sequential chunks associated with the places along the task path, which effectively look ahead h steps. These will be built up by extending the existing chunk of level h forward one step, ie, to the h+1-st step. Thus the height of these chunks equals their span. Let us assume the chunks for the individual states are known, so we are concerned only with the linking chunks. The chunks of length 1 will be the b links from each of the k steps; the chunks of length 2 will be the $b^2$ links from each of the first k - 1 steps (excluding the last which doesn't have two steps to go). And so on. Thus, the number of chunks is:

$$C_{te} = kb + (k-1)b^2 + (k-2)b^3 + ... + (k-h+1)b^h \quad (85)$$

$$C_{te}(h) = f[(k + f - h)\, b^h - (k + f)] \quad \text{where } f = b/(b-1),\ 1 \leq h \leq k \quad (86)$$

The first example indicates in an explicit way how the exponential character arises. The second example, though quite narrow, still suggests that, as particular constraints are taken into account, $C_{te}$ develops polynomial functions of h as well as exponential functions. Thus, it seems reasonable to generalize $C_{te}$ to:

$$C_{te}(h) = S_0(h) + S_1(h)b^h \quad \text{where the S are positive polynomials in h} \quad (87)$$

A critical step in the derivation of the simple model is the identification of $C_{te}(h)$, the chunks to level h provided by the task environment, and C(t), the chunks learned up to a given time. In part, this is built into the structure of chunking. Chunks are learned from the bottom up in terms of existing chunks (C). In part, this comes from the assumption of learning by experience. The stream of chunks of $C_{te}$ presented by the sequence of tasks, trial by trial, becomes the sole source for chunks of C. However, this latter implies that an ergodic assumption is being made, namely, that the time sequence of environments approximates a uniform distribution over the space of all possible environments. Otherwise, no justification would exist for computing $C_{te}$ from the entire space, with all variations assumed equiprobable. We would have to compute the actual set of chunks presented by an analysis of the trial sequence.

In laboratory situations this ergodic assumption may often be satisfied directly by design. Eg, in Seibel's task, trials occurred in blocks of 1023, each block being a random ordering of each of the possible environments. It would be hard to do better. But in nature and in many experiments, especially those with subject control of the trial sequence, significant deviations would no doubt occur.

The ergodic assumption may not be critical. In all cases, $C_{te}$ may be defined as the chunks (of level h) that occur over the actual total trial sequence. This will have some of the same combinatorial properties as the full space, on the assumption that the class of all relevant patterns (ie, chunks) involves some closure operations over the variations of elements, attributes and relations, since what is relevant has to cover the future trial sequence as well as the past. The computation of $C_{te}$ may not be so simple, but the critical question is whether such time cumulation can be approximated by an equation of the form 87, not whether it is a

somewhat more complex instance of such a formula. In any event, we will continue to assume the identification.

### 6.2.3. Learning by Experience

The basic learning rate assumption (chunks are learned at a constant rate) does not seem to require modification. True, the total time hypothesis, which is implied by this assumption, has been found wanting on many experimental tests (eg, Roberts, 1972), which might indicate some more complex formulation. However, these failures take one of two forms. Often error measures (eg, probability correct) are used that are defined on a bounded interval (eg, [0,1]) and hence cannot possibly be linear functions of time indefinitely; this sort of criterion problem is not relevant to our purpose. However, deceleration occurs in situations where relevant learning opportunities become impoverished (eg, failing to learn twice as much about a single nonsense syllable exposed for 8 secs rather than 4 secs). In the present formulation such effects are to be covered by assumptions about what the task environment offers or by assumptions about how experience channels or limits what is learned. Thus, there appears no reason to back off from the assumption that, for the time scales in question (hours to weeks), the human is capable of performing the same basic learning operations in each interval of time, if the environment permits. For this model, the basic learning operation is chunk formation and acquisition.

All that is learned is not necessarily relevant to the task at hand, ie, does not contribute to C. The nub of the assumption that learning is by experience -- and only by experience -- lies in two things: (1) only when attending to aspects of the task relevant to performance does learning occur that is relevant to performance; and (2) attention occurs only when actually performing.

For instance, attention can surely wander and the subject can become distracted. Then no relevant experience is taking place, though the subject may be learning about other things. Such variations find their natural formulation in stochastic models of the learning rate. On average (ie, at the level at which we are modeling) such models still entail that the rate of relevant learning ($\lambda$) is constant, with the effect of distraction showing in a reduced average rate.

However, the present formulation would seem to preclude situations where the subject actively thinks about the task -- inventing and analyzing methods, ruminating about his experience, etc. Thinking per se is not out, as long as it occurs while performing and about the current performance. What seems hard to encompass is study in off-hours, so to speak.

The present generalization does encompass the possibility of a positive asymptote, ie, of an amount of performance that cannot be reduced by any further learning. This implies the subject is engaged in activities that are not relevant to learning about performance. For instance, assume an asymptote of 100 secs per trial, and suppose learning has proceeded to the place were the total performance time (T) were 101 secs. It seems

hardly possible that all the chunks learned throughout the 101 secs provide relevant experience, ie, chunks, for learning on the remaining second. More plausible is that only the experience during the 1 sec (however distributed throughout the total trial) is relevant and contributes chunks to C.

An appropriate formulation of this is that dC/dt proceeds at a rate that is diminished by the requirement to process the asymptote. This is A/T, where A is the asymptote; it effectly smears the processing uniformly throughout the trial.

$$dC/dt = \lambda(1 - A/T) \tag{88}$$

### 6.2.4. The Generalized Chunking Learning Law

Having settled on a more general set of conditions under which to derive a learning law, we recap the equations, using $C = C_{te}$ to obtain the relation between C and the height of chunking in the subject. For convenience we shift from powers of b and c to exponentials:

$$T = P_0(h) + P_1(h)e^{-\gamma h} \quad \text{where the Ps are positive polynomials in h, } \gamma = \log(c) \tag{89}$$

$$C(h) = S_0(h) + S_1(h)e^{\beta h} \quad \text{where the Ss are positive polynomials in h, } \beta = \log(b) \tag{90}$$

$$dC/dt = \lambda(1 - A/T) \tag{91}$$

$$dt/dN = T \tag{92}$$

[NOTE: DO NOT KNOW HOW TO FORMULATE EQUATIONS WITH $C_{te} = C$ EXPLICIT AS AN ADDITIONAL EQUATION, AND $C_{te}$ IN EQUATION 90. THE PROBLEM IS THAT SUCH A SYSTEM PERMITS THE COMBINATION OF EQUATIONS 89 AND 90 TO GIVE $T = F(C_{te})$. BUT THIS CANNOT BE THE CASE WITHOUT USE OF $C_{te} = C$, WHICH IS SITTING AS AN INDEPENDENT EQUATION. SO THE FORMULATION WITH THE h'S IN THE TWO EQUATIONS IDENTIFIED IS NOT RIGHT. BUT DO NO KNOW WHAT IS RIGHT.]

These equations determine T as a function of N, ie, a learning law. Equation 89 determines T as a function of h, $T = F_1(h)$. Equation 90 determines h as a function of C, $h = F_2(C)$, hence T as a function of C, $T = F_1(F_2(C))$. Equation 91 determines C as a function of t and T, which, by means of Equation 92, becomes a function of t and N, $C = F_3(t, N)$; hence, $T = F_1F_2F_3(t, N)$. Equation 92 determines t as a function of N and T, $F_4(N, T)$, yielding $T = F_1F_2F_3(F_4(N, T), N)$. Finally, this last equation can be inverted to yield T as a function of N, $T = F(N)$. Actually, some conditions are necessary on the polynomials involved, $P_0(h)$, $P_1(h)$, $S_0(h)$ and $S_1(h)$. There are monotonicity-like assumptions, which are physically plausible. We did not state them earlier, because they would have seemed unmotivated.

We now show the following theorem:

> Theorem: The system of equations 89 - 92 determines T(N) as an approximate general power low, the approximation being better, the lower the degrees of the polynomials involved.

Let us first determine T as a function of C from Equations 89 and 90:

$$T(C) = P_0(h) + P_1(h)e^{-\gamma h} = P_0(h) + P_1(h)[(C - S_0(h))/S_1(h)]^{-\gamma/\beta}$$

$$= P_0(h) + P_1(h)S_1(h)^{\gamma/\beta}\,[C - S_0(h)]^{-\gamma/\beta} \qquad (93)$$

To eliminate h inside the polynomials, requires expressing h as a function of C. However, before we do that, consider the special case in which the P and S are all constants. Then we have:

$$T(C) = P_0 + P_1 S_1^{\gamma/\beta}\,[C - S_0]^{-\gamma/\beta} \qquad (94)$$

Now Equation 92 can be used to eliminate T in Equation 91 to obtain:

$$dC/dt = \lambda(1 - A(dN/dt)) \qquad (95)$$

This can now be integrated to obtain:

$$C(t, N) = \lambda(t + E_t - AN) \qquad (96)$$

This equation simply expresses that the total learning is the result of the constant rate ($\lambda t$) minus the N periods, each A long, to perform the unlearnable asymptotic processes. $E_t$ is the constant of integration, expressed as the prior time during which experience accrued.

Now, we can eliminate C from Equation 94 to get T(t, N).

$$T(t) = P_0 + \lambda^{-\gamma/\beta}\,P_1\,S_1^{\gamma/\beta}\,[t - AN + E_t - S_0/\lambda]^{-\gamma/\beta} \qquad (97)$$

Equation 97 is close to a power law, but it contains t as well as N. We can eliminate t by forming the differential equation for dT/dN, as in the approach used in the simple version. From Equation 94 we get:

$$dT/dN = -(\gamma/\beta)\,\lambda^{-\gamma/\beta}\,P_1\,S_1^{\gamma/\beta}\,[t - AN + E_t - S_0/\lambda]^{-(\gamma/\beta)\,-\,1}\,(dt/dN - A) \qquad (98)$$

We can use Equation 92 to eliminate dt/dN and Equation 97 inverted to eliminate t. For the latter we have:

$$[t - AN + E_t - S_0/\lambda] = \lambda^{-1}\,P_1^{\beta/\gamma}\,S_1\,[T - P_0]^{-\beta/\gamma} \qquad (99)$$

For comparison with Equation 69 in the simpler version, we reintroduce $\rho = \log(b)/\log(c)$.

$$dT/dN = -(\lambda/\rho)\,P_1^{-\rho}\,S_1^{-1}\,(T - P_0)^{1\,+\,\rho}\,(T - A) \qquad (100)$$

It can be seen from Equation 89 that $P_0$ (being a constant) is the asymptote of T. Thus $A = P_0$ and we get:

$$dT/dN = -(\lambda/\rho)\,P_1^{-\rho}\,S_1^{-1}\,(T - P_0)^{2\,+\,\rho} \qquad (101)$$

Equation 101 is identical with Equation 69, with the substitution of $T - P_0$ for T. Hence, corresponding to Equation 70, Equation 101 integrates to a power law in $T - P_0$, ie, a general power law.

$$T = P_0 + [\lambda(1+\rho)/\rho]^{\,-1/(1+\rho)}\,S_1^{\,1/(1+\rho)}\,P_1^{\,\rho/(1+\rho)}\,(N + E)^{-1/(1+\rho)} \qquad (102)$$

$(E_t - S_0/\lambda)$, having disappeared from Equation 100, does not appear in Equation 102. It shows up in the

initial starting point, E.

Let us now return to the case of general P and S. We need to determine h as a function of C. It will turn out that Equation 94 will be a good approximation to the general situation. From Equation 90 we get:

$$\log(C) = \log(S_0 + S_1 e^{\beta h}) = \beta h + \log(S_1 + S_0 e^{-\beta h}) \tag{103}$$

$$h = (1/\beta)\log(C) - (1/\beta)\log(S_1 + S_0 e^{-\beta h}) \tag{104}$$

Now $S_0$ and $S_1$ are polynomials in h, so we can write a combined polynomial:

$$S(h) = S_1(h) + S_0(h)e^{-\beta h} = (\sum_{i=0}^{m} s_{1,i}h^i) + (\sum_{=0}^{m} s_{0,i}h^i)e^{-\beta h}) = \sum_{i=0}^{m} (s_{1,i} + s_{0,i}e^{-\beta h})h^i$$

$$= \sum_{i=0}^{m} s_i h^i \qquad \text{where } s_i = s_{1,i} + s_{0,1}e^{-\beta h} \tag{105}$$

In the above, m is the degree of the larger polynomial, and the polynomial of lesser degree is simply extended with zero coefficients. The coefficients $s_i$ of S are not independent of h, but they rapidly fade from their initial value of $s_{1,i}+s_{0,i}$ to just $s_{1,i}$. Thus, they can be treated as wobbly coefficients, ie, as coefficients that are indeterminent within a restricted range.

At any time S(h) will behave essentially as some power of h. As h grows, this will be the highest power, $s_m h^m$. However, due to the large size of some coefficients, lower powers may dominate for short intervals at small values of h. Still, the behavior of Equation 105 can be approximated by taking $s_m$ and m as wobbly constants, and writing:

$$h = (1/\beta)\log(C) - (1/\beta)\log(s_m h^m) = (1/\beta)\log(C) - (1/\beta)\log(s_m) - (1/\beta)m\log(h) \tag{106}$$

$$(h + (1/\beta)m\log(h)) = (1/\beta)\log(C) - (1/\beta)\log(s_m) \tag{107}$$

$$h[\beta + m(\log(h)/h)] = \log(C) - \log(s_m) \tag{108}$$

Finally, log(h)/h is a function that varies between 0 and 1/e (.37), and is quite close to .3 between h = 2 and 10. Thus, the expression $[\beta + m(\log(h)/h)]$ is itself just a wobbly constant. Thus we finally get:

$$h = k_1\log(C) - k_2 \qquad \text{where the k are wobbly constants} \tag{109}$$

The use of *wobbly* constants here is short hand for establishing bounds and carrying through arguments with bounds more carefully. We can see later what stipulations need to be made on the wobbly constants that they not invalidate the argument.

From Equation 109 we can now eliminate h from the polynomials $P_0$, $P_1$, $S_0$ and $S_1$, and express them as a function of C. Thus, in Equation 93, T becomes entirely a function of C, hence in Equation 97, entirely a function of t (now taking the polynomials as general). Now, these polynomials depend not on t directly, but on log(t). But log(t) is a very slowly varying function of t compared with t itself, which appears in Equation 97. Thus, to a first approximation, the polynomials are themselves wobbly constants. Hence, our analysis of T(N)

based on the P and S being constants will essentially hold true.

[DO WE NEED TO PUT IN EXPANSION OF P AND S AS POWERS OF LOG(t), SO CAN SEE APPROXIMATION IN TERMS OF $K_j(\log(t))^j(t - K'_j)^{\beta/\gamma}$ ?]

The approximation holds only as t grows, ie, as log(t)/t goes to 0. For small values of t it need not hold at all. The approximation also depends on the degree of the polynomials involved, for it really depends on how rapidly $(\log(t))^m$ / t goes to zero, which is slower the larger is m.

The approximation also depends on some of the polynomials behaving in a reasonable way. The polynomials are already guaranteed to be positive. But if, in particular, $P_1$ and $S_1$ are not reasonably monotonic, then they can nullify the value of t. $P_1$ becoming small permits the polynomial character of $P_0$ to dominate the form of T. $S_1$ becoming small permits the value of $S_0$ to dominate. [SHOULD WE (NEED WE) DERIVE ACTUAL CONDITIONS?]

### 6.2.5. Conclusion

We have now shown that the Chunking Model leads to a power-like law of learning over wide variations in the way chunking enters into performance and the way chunks are generated by the environment. As long as the essential combinatorial aspects remain, the relationship will prevail. Those combinatorial aspects are that chunks cascade in the learner and can be used without total decoding, and that the environment has multiple sources of independent variation. The power-like character depends just on these combinatorial aspects ($\beta$ = log(b) and $\gamma$ = log(c)); details, represented in the model by the polynomials, do not show through over long practice (ie, as t and N increase).

This exercise in generality has been a necessary part of presenting the Chunking Model, since the ubiquity of the phenomena of the power law implies a corresponding generality in the explanation of it.

# 7. CONSEQUENCES AND EXPLANATIONS

Several features of the basic learning data test the model to various degrees. We discuss these first. A substantial amount of work has already been done that bears on the role of chunking, both in performance and in learning. Some of this is pertinent to the present model, and we discuss it next. Finally, there are a few interesting ramifications of the model that are worth discussion, including some implications for the architecture.

## 7.1. The Values of Alpha

Taking the log-log plots at face value produces a range of $\alpha$ that is substantially less than 1. [WAITING FOR THE DATA FROM THE TRANSFORMED GRAPHS]

## 7.2. Convergence: The relation between B and Alpha

We have already noted the *convergence* property of the plots of the log-log law. Formally, this shows up as a positive relationship between B, the initial performance time at N = 1, and $\alpha$, the learning rate: the higher B the larger $\alpha$ (ie, the more rapid the learning). Convergence appears everytime there seems to be a family of learning curves, either ranging over task variation (as in the nine curves in Figure RABBITTCV) or over subjects (as in the eight curves in Figure KOLERS). In some curves the phenomenon of *crossing over* seems imminent. Though we have no data that show it for task families, we do have an example for subject families -- the initially poorer subjects overtaking the initially better (Figure MORAN. This seems acceptable, though possibly interesting. However, crossing over within a family of tasks -- the initially harder task becoming easier -- might seem somewhat alarming. For example, we might find in the Rabbitt, Cummings & Vyas (Figure RABBITTCV) perceptual experiment that searching for more targets in larger displays eventually became easier than fewer targets and smaller displays. As noted in Table DATASUMMARY, this is extrapolated to occur at trial [XX].

Figure BALPHA shows all of the coefficients from the curves we have presented, with the families linked together. [NOTE FEATURES?]

There are two types of explanation for this phenomena. It could be artifactual, resulting from the way the data is displayed or analyzed. Alternatively, though not mutually exclusively, it could represent a psychological phenomena to be explained by theory -- in this case, by the Chunking Model. We will look at both possibilities, starting with the possibilities for artifacts.

### 7.2.1. Do A and E cause convergence

For the simple curve of $BN^{-\alpha}$, B and $\alpha$ are independent, but when the asymptote (A) and the starting point (E) are nonzero, then the measured value of B (T at N = 1) and $\alpha$ (the slope) in log-log space are not independent. Let B* and $\alpha$* be the measured values. Then we have (in part from Section 3.2):

$$T = A + B(N + E)^{\alpha} \tag{110}$$

$$B^* = A + B(1 + E)^{\alpha} \tag{111}$$

$$\alpha^* = \alpha_{inflection} = (\alpha N^* - E) / (N^* + E) \tag{112}$$

$$N^* = N_{inflection} = [BE/\alpha A]^{1/(1+\alpha)} \quad \text{where } E/N^* << \alpha < 1 \tag{113}$$

We are interested in the relation of B* and $\alpha^*$ as the parameters characterizing the curve, namely $\alpha$, A, B and E vary in some way to create a family of curves. It is straightforward to calculate the changes in B* and $\alpha^*$ with changes of the parameters to see whether they shift B* and $\alpha^*$ to create a positive relation, a negative one or some mixture. [JA SAYS HE CAN'T GET THESE FORMULAS]

$$dB^*/dA = 1 \tag{114}$$

$$d\alpha^*/dA = -EN^*/A(N^* + E)^2 \tag{115}$$

$$dB^*/dB = (B^* - A)/B \tag{116}$$

$$d\alpha^*/dB = EN^* / B(N^* + E)^2 \tag{117}$$

$$dB^*/dE = -\alpha(B^* - A)/(1 + E) \tag{118}$$

$$d\alpha^*/dE = (N^* - (1+\alpha)E) / (N^* + E)^2 \tag{119}$$

$$dB^*/d\alpha = -\log(1 + E)(B^* - A) \tag{120}$$

$$d\alpha^*/d\alpha = (N^* - (\log(N^*) + (1-\alpha)/\alpha)E) / (N^* + E)^2 \tag{121}$$

If both derivatives have the same sign, then B* and $\alpha^*$ move together under the variation of a parameter and thus produce a converging set of curves. Conversely, opposite derivatives produce a diverging set of curves. The signs of the derivatives are relatively easy to interpret, if it is noted that $B^*-A \geq 0$ and, though N* rises with E, it does so more slowly than E. [BUT EXPRESSION FOR N* ASSUMES $E/N^* \lambda\tau\lambda\tau \alpha$.] The following statements hold:

- Shifts in A (the asymptote) produce diverging curves.
- Shifts in B (the true initial performance at N = 1) produce converging curves.
- Shifts in E (the starting point) produce diverging curves, as long as $N^* > (1+\alpha)E$. This will be true for small E, but eventually shift for E sufficiently large.
- Shifts in $\alpha$ (the true learning rate) produce diverging curves, as long as N* is greater than $(\log(N^*) + (1-\alpha)/\alpha)$ times E. Again, this will be satisfied for small E, but will shift at some point as E grows.

Thus, it is certainly possible to produce convergence artifactually, if the restrictions for a family are restricted to variation in A, or for large E. However, these conditions do not seem terribly likely to account for the rather universal occurrence of convergence.

One way to test part of these predictions is to ask whether the convergence goes away when optimal E and

A are found. Figure ALPHABKA shows the curves of T−A versus N+E plotted in log-log space. It can be seen that the convergence remains [WHATEVER].

### 7.2.2. Effects of arithmetic averaging

Another possible source of artifact arises from the use of arithmetic averaging. The proper averaging is geometric, since this preserves the power law:

$$T_g = [T_1 T_2 \ldots T_m]^{1/m} = [B(N_1)^{-\alpha} B(N_2)^{-\alpha} \ldots B(N_m)^{-\alpha}]^{1/m} = B[[N_1 N_2 \ldots N_m]^{1/m}]^{-\alpha} \quad (122)$$

$$= B[N_g]^{-\alpha} \quad (123)$$

With arithmetic averaging this does not work. The arithmetic average (eg, $T_a$) is always greater than the geometric average ($T_g$) (unless all values are equal), so we can write:

$$T_a = \tau T_g, \quad N_a = \nu N_g \quad \tau, \nu \geq 1 \quad (124)$$

$$T_a = \tau T_g = \tau B(N_g)^{-\alpha} = B\tau(\nu^{-1} N_a)^{-\alpha} = B(\tau \nu^{\alpha}) N_a^{-\alpha} \quad (125)$$

$$= B^* N_a^{-\alpha} \quad \text{where } B^* = B\tau\nu^{\alpha} \quad (126)$$

$B^*$, the apparent B, now has a dependence on $\alpha$. Since $\nu \geq 1$, this produces a tendency toward convergence -- the larger the $\alpha$, the larger the $B^*$. Only for the STAIR and the Card pointing data do we have the individual measures, to test how bad this is. We find [WHATEVER].

[FURTHER THEORETICAL ASSESSMENT DEPENDS ON GETTING THE FORMS OF $\tau$ and $\nu$. TURNS OUT THIS IS POSSIBLE, BUT NOT YET PUT IN. SHEDS LIGHT ON WHETHER ARITHMETIC AVERAGING CHANGES THE SHAPE OF THE CURVE.]

### 7.2.3. Convergence in the Chunking model

We now turn to substantive explanations for the convergence effect. We start with Equation 74:

$$T = [(1-\alpha)F]^{\alpha} P N^{-\alpha} \quad (127)$$

$$B = [(1-\alpha)F]^{\alpha} P \quad (128)$$

We wish to discuss $B(\alpha)$. Since $B(0) = P$ and $B(1) = 0$, it is clear that B decreases with $\alpha$. However, it actually starts positive and goes through a maximum:

$$dB/d\alpha = B[\log(F) - \alpha/(1-\alpha) - \log(1/(1-\alpha))]$$

$$= B[\log(F) - g(\alpha)] \quad (129)$$

Where we have defined the function $g(\alpha)$ by:

$$g(\alpha) = \alpha/(1-\alpha) + \log(1/(1-\alpha)) = \log[(1-\alpha)^{-1} e^{\alpha/(1-\alpha)}] \quad (130)$$

Since $\alpha \leq 1$ all the individual terms are positive:

$$dB/d\alpha[0] = \log(F) \quad (131)$$

This is positive if $F \geq 1$. Given the interpretation of S, $\lambda$ and P, F is likely to satisfy $F \geq 1$. $F<1$ would imply that the basic environment is very lean compared to the initial performance requirements and the rate of learning. Assuming $F \geq 1$, the question becomes over what range of F does $dB/d\alpha$ stay positive.

At the end, it is straight down:

$$dB/d\alpha[1] = (\log(F) - g(1)) = \log(F) - \infty = -\infty \quad (132)$$

Setting the derivative to be 0 we get:

$$\log(F) = g(\alpha) \quad (133)$$

$$F^* = F_{dB/d\alpha=0} = e^{g(\alpha)} = (1 - \alpha)^{-1} e^{\alpha/(1-\alpha)} \quad (134)$$

As F grows, the range of $\alpha$ over which there is a positive relation between B and $\alpha$ grows. We can plot that range, $0 \leq \alpha \leq \alpha^*$. Actually, the curves are all parallel for different F, simply starting at log(F) and diminishing by $g(\alpha)$, independently of F; so we provide $g(\alpha)$ as well. The initial slope at zero, $dB/d\alpha[0]$, gives a feeling for how strong the postive relationship is. This is just log(F*) which is also $g(\alpha^*)$ by Equation 130.

[HERE'S A TABLE FOR IT, AT THE MOMENT:

| $\alpha^*$ | 0 | .1 | .2 | .3 | .4 | .5 | .6 | .7 | .8 | .9 |
|---|---|---|---|---|---|---|---|---|---|---|
| F* | 1 | 1.24 | 1.61 | 2.19 | 3.25 | 5.44 | 11.20 | 34.37 | 273.0 | 81031 |
| $g(\alpha)$ | 0 | .216 | .473 | .785 | 1.177 | 1.693 | 2.416 | 3.537 | 5.609 | 11.302 |

JUST FOR THE RECORD, ALSO, HERE IS A SIMPLE FORMULA FOR THE SECOND DERIVATIVE.

$$d^2B/d\alpha^2 = B\{[(1/B)dB/d\alpha]^2 - \alpha/(1-\alpha)^2\} \quad (135)$$

### 7.2.4. Do power laws in series cause convergence?

[THERE IS A LAST IDEA -- THAT IF ONE HAD A SERIES OF POWER LAWS, PLOTTING THE T (WHICH IS A SUM, RATHER THAN A PRODUCT) WILL PRODUCE CONVERGENCE IN A MANNER ANALOGOUS TO THE ARTIFACT OF ARITHMETICAL AVERAGING. STILL TO BE WORKED OUT.]

## 7.3. Log-log Learning and Uncertain Choice RT

An important chapter in the analysis of human behavior was the discovery that human choice time appeared to depend on the amount of information required for the decision (Hick, 1952, Hyman, 1953,

Welford, 1968). Figure HYMAN reproduces Hyman's figure (Hyman, 1953) that shows:

$$T = T_0 + T_1 \sum_i p_i \log(p_i)$$

where $p_i$ = probability of occurrence of the i-th stimulus (136)

Hyman used three conditions of uncertainty, the size of the set of equiprobable stimuli, a fixed set of stimuli [TRUE] with varying probability of occurrence, and a fixed set with first order conditional probabilities. All points lie along the same curve. There is evidence for an even stronger form of this law, in which the time of the response to the i-th stimuli is proportional to $\log(p_i)$ individually, not just on average, which is all Equation 136 claims (see Fitts & Posner, 1967, for a review of the evidence).

The work of Seibel, whose data we have used throughout for illustration, was motivated by whether the information theoretic law survived practice or whether the curve flattened out. The choice of a stimulus of 10 bits of uncertainty helped to provide a definitive test, prior work having been limited to a few bits, which is all that can be created by a response with a single finger, rather than chords. Seibel's result, as we have seen but not noted, was that, indeed, learning flattened out the curve so that the information theoretic relation could not be said to survive.

Independently of any precise law of practice or any particular underlying model, the more experience with a stimulus situation the faster should be response to that stimulus. When knowledge of stimulus uncertainty is conveyed by experience (rather than by instruction), there is a covariation of uncertainty with the frequency of occurrence of the corresponding stimulus situation. This implies that response to uncertain stimuli will be slower, solely by virtue of less practice.

Let us compute this effect in the simple situation of k equiprobable stimuli, assuming the power law. In N trials the each stimulus will have been experienced N/k times on average. Thus, the time to respond will be:

$$T = A + B(N/k)^{-\alpha} \quad (137)$$

We wish to look at this equation as a function of k, ie, the amount of uncertainty as in the Hick-Hyman law:

$$T(k) = A + (BN^{-\alpha})\, k^{\alpha} \quad (138)$$

This is certainly not log(k) as required by the information measure. Yet it is not so far from it for suitable values of $\alpha$, as Figure ITPOWER shows. [PLOT OF $k^{\alpha}$ VS k FOR k = 1:8 (IE, H = 0:3). PROBABLY A FAMILY OF PLOTS ALONG WITH log(k)] I BELIEVE BY THE WAY THAT THERE IS A DISCUSSION SOMEWHERE OF sqrt(k) AS A FORM FOR THE LAW -- SEEM TO REMEMBER WELFORD, BUT DON'T FIND IT ANYWHERE.]

For stimuli of differential probability of occurrence one gets the same curve with $1/p_i$ instead of k. The

expected choice RT, corresponding to equation 136 and Figure HYMAN would then be:

$$T = A + (BN^{-\alpha}) \sum_i p_i(1/p_i)^{\alpha} \qquad (139)$$

Again, this is a good approximation over small ranges of bits.

[WHAT ABOUT SEQUENTIAL PROBABILITIES?]

It can be seen that the main implication does not depend on on any specific model or even law. Use of the exponential would also lead to a law. However, it would be $e^{-\alpha N/k}$ whose behavior asymptotes and which provides a substantially poorer approximation. In one respect, the Chunking Model specifically adds to the credibility of this result. Namely, it supports the assumption that each of the stimuli is to be seen as a distinct learning situation which can be treated in isolation from each other. For the model claims that, as chunking proceeds, the subject arrives at characterizations of the environment that are increasingly narrow, and thus define separated learning situations, to which the analysis above applies.

We do not suggest that this learning mechanism is all that is behind choice RT under uncertainty. However, it does follow from the theory that the effects above must be present in the choice RT data, possibly mixed in with the effects of other mechanisms. The results of Seibel, of course, show that to be true.

### 7.4. Relation to Existing Work on Chunking

Much of the existing work on chunking has focussed 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 (DeGroot, 1965, Chase & Simon, 1973) bears directly on the present model. The basic phenomena 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 five second exposure, while masters can recall most of the pieces.

A well articulated theory has evolved to explain this striking phenomena. 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 position 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 model 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 Figure STAIR can be taken as a reasonable analogue of the chess game.

[ARE THERE OTHER CONNECTIONS?]

### 7.5. The Structure of the Task Environment

An important, and indeed pleasing, feature of the Chunking Model is that it implies a connection between the structure of the task environment and the learning behavior of the subject. The richer the task environment -- ie, ensemble of environments with which the subject must potentially cope -- the more difficult his learning.

The theory makes some quite simple assertions about this richness: it is measured by a single parameter, $\beta$ = log(b). In fact, the theory seems determined to be a little too simple minded here. For we saw in the generalized version, that the same factor came through. Thus, it may not be easy to fashion versions of the theory that will depart seriously from this. This arises, as we noted earlier, because the combinatorial growth tends ultimately to dominate all other sources of variation. This very feature, while it throws the theory into jeopardy by seemingly being incapable of generalization, also accounts for the ubiquity of log-log linear learning.

[MORE]

### 7.6. Transfer of Training

### 7.7. When does Learning Not Occur?

Learning does not always occur, log-log linear or otherwise. What does the theory have to say about these? A good example comes from some work by Kristofferson (197x) (Kris7x) on the basic Sternberg binary classification paradigm.

[MORE]

## 7.8. Implications for the Cognitive Architecture

One of our original motivations in looking into the power law was to see if it could provide an additional general constraint or two on the nature of the architecture that supports information processing in the human. In line with our observation in the introduction about the types of explanations that lie behind ubiquitous phenomena, we did not expect other than extremely broad general features. Far from discouraging us, that is exactly what we wanted as an addition to the set of design considerations that would constrain the search for an acceptable architecture.

[MORE]

# 8. 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, hence relevant to the modern work in automatization. The commitment to write this paper 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 79, 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 exponent -1, ie, 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 AN 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 (Figures ANDERSON and NEVESANDERSON) that bolstered this.

This picture seemed reasonably satisfactory, though the existence of log-log linear industrial learning curves (Figure HIRSCH) nagged a bit, as did the persistence of some of our colleagues in believing in the argument of mixtures, despite our theoretical arguments (see Section 5.1) to the contrary. However, as we proceeded to write the paper, 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 paper.

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 (Figures SNODDY, SEIBEL), as due to special causes, distinct from their description as a generalized power function.

3. The data does not fit the simple power law (ie, 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. There is no way to confirm either whether the data (1) fits within the hyperbolic subfamily or (2)

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 phenomena is the ubiquity of the learning data, ie, 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 (ie, industrial) situations yield similar curves. Our restriction was dictated purely by the momentary need to bound the research effort.

6. An (apparently) important phenomena is that the learning rate parameter ($\alpha$) clusters well below $\alpha = 1$, in general around $\alpha = .2 - .3$. However, transformation to the appropriate general coordinate system (N + E, T - A), with arbitrary starting point and asymptote, raises these rates until they are around 1. Though it remains possible to maintain that $\alpha \leq 1$, that generalization does not seem especially safe. [IF IT REMAINS SO]

7. An important phenomena is the *convergence* of experimentally related families of curves. Though artifactual sources could possibly contribute to the phenomena (general A and E in special cases, and arithmetic averaging), the phenomena appears to be genuine and to require explanation.

   The subphenomena of *crossing* in task families still remains ephemeral.

8. Psychological models that yield the power law with arbitrary rate ($\alpha$) are difficult to find. (Positive asymptotes and arbitrary starting points are, of course, immediately plausible, indeed, unavoidable.) The Chunking Model is the only satisfactory one we have been able to develop.

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

10. The various models are not mutually exclusive, but provide an array of sources of the power law. Several hyperbolic mechanisms could co-exist 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 a mixing to produce hyperbolic learning curves.

11. An attractive (and maintainable) option is that the entire phenomenon is do 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.

    The convergence would should probably be assigned to artifact, though it is not clear such a move is satisfactory.

12. 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.

13. [MORE]

## 9. REFERENCES


Blackburn, J. M. *Acquisition of Skill: An analysis of learning curves.* Technical Report, I.H.R.B., 1936.

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

Calfee, R. C. *Human Experimental Psychology.* New York: Holt, Rinehart and Winston 1975.

Card, S. K., English, W. K. & Burr, B. Evaluation of mounst, 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.

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).

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 hypothosis 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.: Erlbaum 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. & Posner, M. I. *Human Performance.* Belmont, CA: Brooks/Cole 1967.

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 Melton, A. W. (Ed.), *Categories of Human Learning*, New York: Academic Press, 1964.

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 Melton, A. W. & Martin, E (Ed.), *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. Acquistion of automatic processing in perceptual and associative learning. In Rabbitt, P. A. M. & Dornic, S. (Ed.), *Attention and Performance V,* New York: Academic Press, 1974.

Lindsay, P. & Norman, D. *Human Information Processing: An introduction to psychology, 2nd ed.* New York: Academic 1977.

Lippert, S. Accounting for prior practice in skill acquisition studies. *International Journal of Production Research*, 1976, 14, 285-293.

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.

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

Newell, A. Harpy, production systems and human cognition. In Cole, R. (Ed.), *Perception and Production of Fluent Speech*, Hillsdale, N.J.: Erlbaum, 1980.

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

Rabbitt, P., Cumming, G. & Vyas, S. Improvement, learning and retention of skill at visual search. *Quarterly Journal of Experimental Psychology*, 1979, 31, 441-459.

Reisberg, D., Baron, J. & Kemler, D. G. Overcoming Stroop interference: The effects of practice on distactor 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.

Roberts, W. A. Free recall of word lists varying in length and rate of presentation: A test of the total-time hypothosis. *Journal of Experimental Psychology*, 1972, 92, 365-372.

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

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

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.

Steven, 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.

## 10. REFERENCE NOTES

1. Anderson, J. Private communication, 1980.

2. Lewis, C. Speed and Practice, undated.

3. Moran, T. Private communication, 1980.

4. Neves, D. & Anderson, J. Private communication, 1980.

# Table of Contents