Psychonomic Bulletin & Review
2000, 7 (2), 185-207

The power law repealed: The case for an exponential law of practice

ANDREW HEATHCOTE and SCOTT BROWN
University of Newcastle, Callaghan, Australia

and

D. J. K. MEWHORT
Queen's University, Kingston, Ontario, Canada

The power function is treated as the law relating response time to practice trials. However, the evidence for a power law is flawed, because it is based on averaged data. We report a survey that assessed the form of the practice function for individual learners and learning conditions in paradigms that have shaped theories of skill acquisition. We fit power and exponential functions to 40 sets of data representing 7,910 learning series from 475 subjects in 24 experiments. The exponential function fit better than the power function in all the unaveraged data sets. Averaging produced a bias in favor of the power function. A new practice function based on the exponential, the APEX function, fit better than a power function with an extra, preexperimental practice parameter. Clearly, the best candidate for the law of practice is the exponential or APEX function, not the generally accepted power function. The theoretical implications are discussed.

> Curve fitting without benefit of a model is notoriously a black art.
>
> —Newell and Rosenbloom (1981, p. 23)

The benefits from practice follow a nonlinear function: Improvement is rapid at first but decreases as the practitioner becomes more skilled (see, e.g., Thorndike, 1913). The idea that a simple nonlinear function might describe practice effects in a broad range of tasks was championed by Newell and Rosenbloom's (1981) influential chapter entitled "Mechanisms of Skill Acquisition and the Law of Practice." The "law of practice" in the title concerns the relationship between response time (RT) and number of practice trials. Newell and Rosenbloom examined data from a wide range of tasks. When they compared power and exponential functions as possible forms for the law of practice, power functions provided better fits than exponential functions in every case.

The power function is now treated as the *law* of practice. In J. R. Anderson's (1982) words, "one aspect of skill acquisition . . . distinguished . . . by its ubiquity . . . is the log-linear or power law for practice" (p. 397). A decade later, Logan (1992) echoed the same conviction: "The power law is ubiquitous. It occurs in virtually every speeded task" (p. 883). In accord with its status as a law, most research subsequent to the publication of Newell and Rosenbloom's (1981) findings has assumed a power function, rather than testing to determine whether it provides a better description than other functions (e.g., Cohen, Dunbar, & McClelland, 1990; Kramer, Strayer, & Buckley, 1990; Logan, 1988, 1992), or has assumed that a power function holds for each component of performance (e.g., Delaney, Reder, Staszewski, & Ritter, 1998; Rickard, 1997).

The power function's status as a law has also made it a gold standard by which to judge the success of models of skilled performance, including ACT and related models (J. R. Anderson, 1982; J. R. Anderson & Schooler, 1991), the component power laws model (Rickard, 1997), network models (MacKay, 1982; Cohen et al., 1990), instance theories (Logan, 1988, 1992; Nosofsky & Palmeri, 1997), and Newell and Rosenbloom's (1981) chunking model (see also Rosenbloom & Newell, 1987a, 1987b). Logan (1988) leaves no doubt about the importance of the form of the practice function for theories of skill acquisition: "The power-function speedup [is] a benchmark prediction that theories of skill acquisition must make to be serious contenders" (p. 495; see also Cohen et al., 1990, and Palmeri, 1997, for similar views).

However, we contend that the evidence supporting a power law of practice is flawed. Although theories of skill acquisition model learning in individuals, the bulk of the evidence favoring the power law is based on fits to averaged data. There is little empirical evidence from *individual learners* for *individual learning conditions* that a power function describes skill acquisition better than does

We acknowledge financial assistance from Australian Research Council Grant 42/280/022 to R. Heath and A.H., a University of Newcastle Research Management Committee grant to A.H., and grants from the Natural Science and Engineering Research Council of Canada to D.J.K.M. Thanks to everyone who contributed data to the survey and to reviewers of a previous version of this paper, including J. R. Anderson, W. Estes, G. Logan, R. D. Luce, T. Rickard, and J. Wixted. Correspondence concerning this article should be addressed to A. Heathcote, Department of Psychology, University of Newcastle, Callaghan, 2308, Australia (e-mail: heathcote@psychology.newcastle.edu.au).

an exponential function. Data from all but one of the tasks examined by Newell and Rosenbloom (1981), for example, were averaged over subjects, conditions, or practice blocks. In the few published comparisons that report analysis of data from individual subjects, the exponential function fit better than the power function (Josephs, Silvera, & Giesler, 1996; Rosenbloom & Newell, 1987b).

The mismatch between theory and evidence is more than a minor technicality: It has been known from the psychological literature for almost 50 years that average curves need not take the same form as the individual curves making up the average (e.g., Estes, 1956; Kling, 1971; Sidman, 1952). Hence, the form of the average practice function does not unambiguously indicate the form of the components of the average. Moreover, recent work shows that linear averaging yields a composite that is systematically biased toward the power function, when compared with the exponential function (e.g., R. B. Anderson & Tweney, 1997; Myung, Kim, & Pitt, in press). Hence, evidence once thought to favor the power law may be artifactual.

The form of the practice law may also seem to be an unsolvable technical issue, rather than an important psychological question. Estes (personal communication, May 1997) has indicated that the form of the practice law does not constrain theory enough. In his words, a "generation of budding learning theorists (Bower, Bush, Estes, Greeno, Hunt, Restle) produced mountains of analyses showing how easily the forms of particular performance curves can be mimicked by many alternative models."

Nevertheless, the form of the practice law does carry an important implication about the nature of learning. As we will note more fully later, an exponential function implies a constant learning rate relative to the amount left to be learned. By contrast, the power function implies "a learning process in which some mechanism is slowing down the rate of learning" (Newell & Rosenbloom, 1981, p. 18). The question at issue, then, is whether the slowing implied by the power function is part of skill acquisition. Repealing the power law of practice in favor of an exponential (or other) law has serious implications for all theories of skill acquisition—especially for those developed in order to account for the power law.

In light of the ambiguity concerning empirical support for the power law, we report the results of a survey that systematically assessed the form of the practice function for individual learners and learning conditions in paradigms that have shaped theories of skill acquisition (see Table 1 for a summary of the paradigms and data sets). In the next section, we review the properties of candidate practice functions and propose a new practice function, the APEX function, that expedites our analysis. Subsequent sections describe our method and results. The results can be easily apprehended from Table 2 (which tabulates the average fit of the candidate practice functions in each data set) and from Figures 1, 2, and 3 (which present the percentage of cases in each data set that are best fit by a candidate practice function). Finally, the Discussion section examines the implications of our results for the measurement of practice functions and for theories of skill acquisition.

PRACTICE FUNCTIONS

Equations 1 and 2 are the power and exponential functions used by Newell and Rosenbloom (1981) to fit practice data. $E(\mathbf{RT}_N)$ is the expected value of **RT** on practice trial N. The boldface notation indicates that **RT** is a random variable. When referring to observed response times (i.e., to samples from **RT**), we will use the notation RT. A_P and A_E are the expected values of **RT** after learning has been completed for the power and the exponential functions, respectively. An asymptote parameter is necessary when modeling response time, even in a highly skilled subject, because performance is limited by physical constraints, such as neural integration time and motor response time. B_P and B_E are the change in the expected value of **RT** from the beginning of learning ($N + E = 1$ for the power function, or $N = 0$ for the exponential function) to the end of learning (the asymptote). Hence, B_P and B_E indicate the range over which practice speeds responding:

$$E(\mathbf{RT}_N)_\mathrm{P} = A_\mathrm{P} + B_\mathrm{P}(N + E)^{-\beta}, \qquad (1)$$

and

$$E(\mathbf{RT}_N)_\mathrm{E} = A_\mathrm{E} + B_\mathrm{E}e^{-\alpha N}. \qquad (2)$$

The amount of nonlinearity displayed by the practice function is controlled by its rate parameter: α for the exponential function and β for the power function. The power function has one extra nonlinear parameter: E. It represents the subject's prior learning from practice before experimental measurement (the rationale for the E parameter was first suggested by Seibel, 1963). Most subsequent researchers, however, have fit the simpler three-parameter version of the power function with E fixed at zero (e.g., Logan, 1988, 1992).

An extension of the exponential function to include a parameter corresponding to E is redundant, because the exponential function is translation invariant—that is,

$$B'_\mathrm{E}e^{-\alpha(N+E)} = B'_\mathrm{E}e^{-\alpha E}e^{-\alpha N} = B_\mathrm{E}e^{-\alpha N},$$

where $B_\mathrm{E} = B'_\mathrm{E}e^{-\alpha E}$. Hence, the effect of prior practice is incorporated into the estimate of B_E for the three-parameter exponential function.

Newell and Rosenbloom's (1981) evidence for a power law of practice was based on a comparison of the fit of the three-parameter exponential function with the fit of the four-parameter power function, which they call the general power function. We will adopt the terms *general power function* and *power function* when referring to the four- and three-parameter (E fixed at zero) versions of Equation 1, respectively.

Apart from the difficulties introduced by averaging, Newell and Rosenbloom's (1981) analysis is open to criticism on two technical grounds.

1. The first technical criticism concerns the number of parameters in the equations considered by Newell and

Table 1
Summary of Data Sets Fit in the Survey

Source	Name	N	Length	Ss	Errors	Censor (msec)
Strayer and Kramer (1994b, 1994c)	MS1	36	606–711	6	Increase*	150<RT<1,500
	MS2	192	654–716	32	Increase	150<RT<1,500
Strayer and Kramer (1994a)	MS3	132	625–860	22	Decrease	150<RT<1,500
Palmeri (1997)	Count1	120	175–208	4	Decrease*	None
	Count2	288	125–160	4	Decrease*	None
	Count3	360	129–160	5	Decrease*	None
Rickard and Bourne (1996)	Math1	384	63–90	24	Decrease*	200<RT<5,000
Rickard (1997)	Math2	228	62–90	24	Decrease*	None
	Math2a	125	>9	-	-	None
	Math2m	228	>9	-	-	None
Reder and Ritter (1992)	Math3	79	8–20	20	NA	None
	Math3a	44	>7	-	-	None
	Math3m	38	>7	-	-	None
	Math4	63	8–20	16	NA	200<RT<18,000
	Math4a	50	>7	-	-	200<RT<18,000
	Math4m	14	>7	-	-	200<RT<18,000
Schunn, Reder, Nhouyvanisvong, Richards, and Stroffolino (1997)	Math5	65	8–28	22	NA	None
	Math5a	57	>7	-	-	None
	Math5m	35	>7	-	-	None
Rickard (1997)	AA1	504	25–84	21	Decrease	200<RT<10,000
	AA1a	157	>9	-	-	200<RT<10,000
	AA1m	489	>9	-	-	200<RT<10,000
Smith and Mewhort (1994)	AA2	288	80	24	Decrease*	None
Heathcote and Mewhort (1993)	VS1	192	200	24	Decrease*	None
Carrasco, Ponte, Rechea, and Samperdo (1998)	VS2	120	63–88	10	Increase	None
Heathcote and Mewhort (1993)	VS3	128	160	8	Decrease	None
Verwey (1996)	Key1t	72	45–613	36	Increase*	None
	Key1c	180	45–613	-	-	None
	Key1k	648	45–613	-	-	None
	Key2t	72	67–1,353	36	Flat	None
	Key2c	180	67–1,353	-	-	None
	Key2k	648	67–1,353	-	-	None
Brown and Heathcote (1997)	Key3	56	228–300	4	NA	200<RT<10,000
J. R. Anderson, Rincham, and Douglass (1997)	Rule1	26	32	26	Decrease*	None
	Rule2	88	32	22	Decrease	None
	Rule3	180	32	45	Decrease*	None
Kail and Park (1990)	MR1c	96	35	8	NA	None
	MR1a	96	35	8	NA	None
Ringland and Heathcote (1998)	MR2c	576	18–32	12	Decrease	None
	MR2a	576	18–32	12	Decrease	None

Note—N, the number of practice series in the data set; Length, the range of lengths for practice series in each data set; Ss, the number of subjects in each data set; Errors, the results of tests on the main effect of practice on accuracy, with a * indicating significance at the 95% confidence level; where results were not significant trends are indicated; Censor, criteria used, if any, to censor outliers from the data set.

Rosenbloom (1981). They compared the three-parameter exponential function against the four-parameter general power function. One might think that the latter function would be more flexible, since it has an extra parameter. In particular, the extra flexibility may allow a general power function to mimic an exponential function, making the general power function almost impossible to falsify by a comparison with the simpler exponential.

The relationship between the general power and the exponential functions is illuminated by expressing them as differential equations:

$$\frac{\partial\left[E(\mathbf{RT}_N)_\mathrm{P}\right]}{\partial N} = \frac{-\beta}{N+E}\left[E(\mathbf{RT}_N) - A_\mathrm{P}\right], \qquad (3)$$

and

$$\frac{\partial\left[E(\mathbf{RT}_N)_\mathrm{E}\right]}{\partial N} = -\alpha\left[E(\mathbf{RT}_N) - A_\mathrm{E}\right]. \qquad (4)$$

Equations 3 and 4 can be compared by using their *relative learning rate* (RLR), defined as minus the rate of change of expected **RT** divided by the amount left to be learned (i.e., RLR equals the multipliers of the bracketed terms on the right side of Equations 3 and 4). The defining characteristic of an exponential function is a constant RLR (α) at all levels of practice. For the general power function, by contrast, RLR is a hyperbolically decreasing function of practice trials $[\beta/(N+E)]$.

Strictly speaking, the general power function can exactly mimic an exponential function only when $\alpha = 0$ (i.e., a flat function). However, for practice series of finite

Table 2
Results for Published and Unaveraged Data Sets

	R^2				$\%\hat{A} < 150$ msec				% Significant			
Data Set	E	P	APEX	GP	E	P	APEX	GP	Overall	$\hat{\alpha}'$	$\hat{\beta}'$	APEX
MS1	.051	.047	.054	.053	13.9	80.1	47.2	44.4	97.2	41.7	8.3	2.8
MS2	.124	.119	.130	.128	11.5	61.5	24.0	31.3	93.2	48.4	27.1	5.7
MS3	.055	.048	.056	.055	19.0	72.0	39.4	37.1	81.1	50.0	9.1	0.0
Count1	.647	.507	.651	.637	18.3	91.7	23.3	30.0	100.0	95.0	20.8	16.7
Count2	.607	.472	.613	.601	11.1	88.9	18.1	30.2	100.0	91.3	13.5	8.0
Count3	.592	.489	.599	.590	5.0	88.3	14.7	25.8	100.0	86.4	18.6	11.7
Math1	.166	.154	.173	.166	19.0	39.1	21.6	21.1	47.9	16.7	9.1	0.3
Math2	.658	.577	.679	.669	19.2	45.7	48.7	54.4	99.6	82.5	32.0	21.5
Math3	.394	.377	.419	.411	36.6	51.2	41.7	38.0	40.5	6.3	2.5	0.0
Math4	.236	.219	.253	.251	23.4	45.3	28.6	38.1	25.4	6.3	3.2	0.0
Math5	.540	.485	.546	.530	33.7	51.1	49.2	49.2	69.2	30.8	0.0	0.0
AA1	.619	.497	.628	.615	18.1	64.0	30.0	44.8	99.4	84.9	11.1	5.6
VS1	.277	.221	.282	.280	19.8	93.2	31.3	48.4	98.4	81.3	10.4	5.2
VS2	.261	.218	.265	.262	7.8	51.6	24.2	29.2	100.0	96.4	17.9	14.3
VS3	.305	.265	.313	.312	7.0	86.7	14.1	28.9	97.7	68.8	13.3	3.9
Key1t	.624	.602	.633	.624	1.4	52.8	8.3	13.9	95.8	52.8	36.1	36.1
Key2t	.185	.176	.187	.179	45.8	63.9	56.9	52.8	88.9	31.9	6.9	1.4

Note—E, P, APEX, and GP refer to the exponential, power, APEX, and general power functions, respectively. Significance is assessed at the 95% confidence level.

length, large values of E can make the RLR almost constant, and any value of α can be approximated, particularly smaller values. In experimental measurement, practice trials vary over a limited range. If the estimate of E is much larger than the range, the effective RLR for the general power function is a constant, approximately equal to β / E. Consequently, the general power function can mimic exponential data, using large estimates of E, although such fits are invariably poorly behaved, because the parameter estimates are highly correlated—that is, large estimates of E are associated with large estimates of β and, particularly, of B_{P}.

Newell and Rosenbloom (1981) noted the association between large estimates of E and large estimates of B_{P} when they fit the general power function to simulated exponential data. We reported the same behavior in fits to several data sets from a visual search paradigm (Heathcote, 1990; Heathcote & Mewhort, 1995). When E is large, large values of β will occur, to allow the general power function to approximate a constant RLR greater than zero. Very large values of B_{P} allow the general power function to approximate the decrease in the expected value of **RT** from the beginning to the end of measurement. This counteracts the tendency of a general power function to become flat when it mimics exponential data.

2. The second technical criticism concerns the way in which Newell and Rosenbloom (1981) fit practice functions. To save on computation, they fit by minimizing squared deviations in $\log(\mathrm{RT} - \hat{A}_{\mathrm{P}})$ and $\log(\mathrm{RT} - \hat{A}_{\mathrm{E}})$. Given that A_{P} and A_{E} estimate *expected values*, it would not be unusual to observe samples from **RT** that are less than these expected values; in such cases, the measures $\log(\mathrm{RT} - \hat{A}_{\mathrm{P}})$ and $\log(\mathrm{RT} - \hat{A}_{\mathrm{E}})$ are undefined. It is difficult to know how to deal with undefined values without biasing or distorting fits (Newell & Rosenbloom did not describe their approach). Undefined values will not occur, of course, if near-asymptotic performance is not measured or if the variability of **RT** shrinks to zero with practice. The latter condition is unlikely, since, even in very fast and simple tasks, RT remains variable (see Luce, 1986, for numerous examples). Such problems are likely to be more pronounced in individual-subject data, since they are noisier than averaged data.

To check whether the fitting method biased their results, we obtained a subset of the data in Newell and Rosenbloom's (1981) survey. We refit their data by minimizing squared deviations in RT (i.e., the generally accepted method of ordinary least squares) instead of $\log(\mathrm{RT} - \hat{A}_{\mathrm{P}})$ and $\log(\mathrm{RT} - \hat{A}_{\mathrm{E}})$. For the averaged data, the power function still fit better than the exponential function. However, for the two data series that were not averaged (times to win and to lose the Stair card game), the exponential function provided a better fit than the power function.[1] In other words, the only evidence favoring the power function for unaveraged data in Newell and Rosenbloom's survey turns on the adequacy of their fitting method.

The two technical criticisms not only call Newell and Rosenbloom's (1981) findings into question[2] but also raise a dilemma. The general power function is clearly a plausible extension of the power function, at least when estimates of E are reasonable, but it cannot be fairly compared directly with the exponential function. To escape the dilemma, we propose a new[3] four-parameter practice function, the APEX function, which nests (i.e., contains as special cases) both the power and the exponential functions:

$$E(\mathbf{RT}_N) = A + Be^{-\alpha' N} N^{-\beta'}. \tag{5}$$

The APEX function has an RLR that is the sum of the RLRs for the power and exponential functions: $\alpha' + \beta' / N$.

Consequently, its RLR decreases like a power function early in practice but approaches an asymptotic value potentially greater than zero (α') later in practice.

We propose the APEX function for two reasons.

1. The APEX function nests both the power and the exponential functions. Hence, fitting the APEX function to exponential data will likely be better behaved than fitting the general power function to exponential data. Because the APEX function can exactly fit both power and exponential data, it can be used to adjudicate between the two alternatives. If the APEX function provides a better fit than the power function, an exponential component is supported. If the APEX function provides a better fit than the exponential function, a power component is supported. If both conditions apply, the full APEX function is supported.

Tests comparing the fit of APEX, power, and exponential functions can be applied to individual subjects and conditions. This allows for the possibility that the form of the practice function might differ between individuals or conditions. The significance of any improvement in fit can be assessed *for each individual subject and condition* by a straightforward nested-model change-of-R^2 test. Nested-model tests have an important advantage when dealing with nonlinear models: They need assume only a linear approximation to the intrinsic curvature of the function and are not affected by nonlinear effects of changes in parameters (see Bates & Watts, 1988).

2. Comparison of the fit of APEX and general power functions not only provides a fair test but also illuminates a theoretically crucial point. Many models of skill acquisition acknowledge that performance may be the result of a sum of component processes (e.g., Kirsner & Speelman, 1996) or the result of a mixture of processes, such as algorithmic and memory-based processing (e.g., Logan, 1988, 1992; Rickard, 1997).

In the case of a sum, processes with a relatively faster learning rate will approach their asymptote quickly and then will cease to affect the rate of change of RT; as a result, asymptotic learning will be dominated by the slower learning processes. The transition should be evident as a decrease in the RLR early in practice, even if all component processes are exponential (i.e., even if all have constant RLRs). Asymptotic learning rates, however, will reveal the true RLR function of (at least) the slow component processes. A superior fit for the APEX function supports an asymptotic RLR greater than zero, a finding inconsistent with component power functions. A superior fit for the general power function supports an asymptotic RLR that approaches zero, a finding inconsistent with component exponential functions.

A similar argument can be made for a mixture of memory and algorithmic processes. Even if each component learns exponentially, RLR can decrease early in practice, because of the transition from trials controlled by slow algorithmic processing to trials controlled by fast memory-based processing. After sufficient practice, however, learning should be dominated by the memory-based process. If memory-based processing follows a power function, the asymptotic RLR will approach zero, and the general power function will provide a superior fit to the APEX function. If memory-based processing follows an exponential function, the APEX function will provide the superior fit, with an estimate of α greater than zero.

Several experiments in the survey required subjects to report their processing strategy as algorithmic or memory based. Hence, we were also able to test the function describing each component explicitly in these experiments (assuming, of course, that the subjects' self-reports accurately describe their processing strategies).

METHOD

Analysis Techniques

We fit power, exponential, general power, and APEX functions, using ordinary least squares minimization on correct trial data. The fits were obtained by using a Simplex search algorithm (Press, Flannery, Teukolsky, & Vetterling, 1986), because it is robust when fitting is ill conditioned. Fits were validated by comparison with the outputs of nonlinear regression programs provided in the SPSS and S+ statistical packages.

When data are exponential, fits of the general power function tend to be ill conditioned. In particular, fitting is problematic, because estimates of B_P diverge to very large values that cannot be represented with accuracy, even in double precision. To avoid the problem, estimates of B_P were constrained to be less than 10^{10} msec (around 115 days). Estimates of the asymptote parameters were constrained to be greater than zero for all fits, in order to ensure that the estimates were plausible.

To ensure that the best fits of the general power function were found, multiple fits were performed with starting estimates for E equal to 1%, 10%, 50%, and 100% of the length of the series. The best fit was then selected and compared with the fit of the power function for the same series. Where the general power function fit was worse than the power function fit, as sometimes happened owing to correlated parameters, the power fit was substituted for the general power fit. Ill-conditioned power, exponential, and APEX fits almost never occurred, and starting points for fitting were easy to obtain by heuristics. Where the fit of an APEX function was worse than the fit of either the power or the exponential function for the same series, refits were performed by using starting values close to both the exponential and the power solutions. The best fit among the refits and the power and exponential fits was selected.

Experimenters often censor their data to delete outliers. Whenever the experimenter censored his or her data, we followed the same procedure. If an experimenter did not report censoring, we removed obvious outliers (see Table 1 for the criteria used). Across all the data sets, however, very few observations were censored.

Because RT can vary as a function of accuracy, we also calculated the main effect of practice on errors for each block of trials. Most frequently, errors decreased with practice, but they also increased or remain unaffected by practice in some data sets. Hence, the results for RT are not correlated in any simple pattern with changes in accuracy, at least across different data sets.

All the fits used learning series broken down by subjects and by within-subjects factors or learning examples. Where strategy reports were available, the learning series were also broken down by strategy for supplementary analyses. In some cases, data sets broken down by strategy produced series that were too short to obtain reliable fits. Such data sets were excluded. The data for production of key sequences were divided into trials on the 1st and subsequent days, because new instructions that clearly influenced learning were given on the 2nd day of practice (see Verwey, 1996, Figure 3,

p. 548). Exponential, power, APEX, and general power functions were then fit to each learning series.

Several approaches were used to compare the overall performance of the different functions. The proportion of learning series for which an exponential function provided a better fit than a power function (on the basis of R^2) was tallied for each data set (see Figure 1). A similar comparison was made between the four-parameter functions (APEX and general power; see Figure 2). Binomial confidence intervals were used to determine whether preference for either member of each pair of functions was significant in each data set. For all functions, the average value of R^2 for each data set was calculated and compared (see Table 2). In addition, the improvement in fit of the APEX function, relative to both the exponential and the power functions, was used to provide nested-model tests. Table 2 reports the percentage of learning series for each data set where significant results for the nested-model tests supported an exponential and/or a power component.

Data Sets

We fit 40 sets of data; collectively, the data represent 7,910 learning series from 475 subjects in 24 experiments taken from 13 published and 3 unpublished sources. The unpublished data (Brown & Heathcote, 1997; Ringland & Heathcote, 1998; Smith & Mewhort, 1994) were collected in our laboratories and were analyzed to clarify and expand on results from the published data sets.

Table 1 summarizes the characteristics of the data sets used in the survey. Each data set was given a unique acronym used to index the summaries of results. For data sets that were broken down by strategy, the "Length" column indicates the criterion length used to exclude short series. Table 1 also reports the results of tests on the effect of practice on error rates. The following sections describe both the paradigms from which the data were drawn and the experimental factors used to produce separate series for each data set.

Memory search. In the memory search tasks, subjects studied a list of words and then were asked to indicate whether or not a probe word had appeared in the study list. The words used in the list were selected to represent particular semantic categories, and semantic category was mapped consistently to the word's use as a target or a distractor item. Fits used series broken down by subjects and within-subjects factors.

The data in the MS1 set are the consistently mapped trials from mixed consistent/varied mapping training blocks from Experiment 2 of Strayer and Kramer (1994b). The data indexed by the label MS2 refer to the consistently mapped training blocks from Experiment 2 of Strayer and Kramer (1994b), from Experiments 4, 6, and 7 of Strayer and Kramer (1994c), and from an unpublished two-alternative forced-choice version of the task. The experiments in the MS2 data set came from very similar paradigms and individually produced the same pattern of results as the overall data set, so they were grouped together. For both MS1 and MS2, two factors were manipulated within subjects: target/distractor probe and memory load (two, four, or six items).

The data in the MS3 set are consistently mapped trials from young subjects (ages, 18–21) from Strayer and Kramer (1994a). Two factors were manipulated within subjects: target/distractor probe and memory load (two, four, or six items), and one factor was manipulated between subjects: speed versus accuracy instructions.

Counting. In the counting tasks, subjects were shown different patterns of 6–11 dots and a spelled-out number; they were asked verify whether the number of dots in the pattern matched the spelled-out number. All the data were taken from Palmeri (1997). Each experiment used a number of unique patterns, and fits included series from each pattern. Fits used data broken down by subjects and dot pattern.

The data in the Count1 set are the training series from Experiment 1. The number of dots was manipulated within subjects. There were 30 patterns, with 5 patterns of array size.

The data in the Count2 set are the training series from Experiment 2. The number of dots and the similarity of dot patterns (none, low, and moderate) were manipulated within subjects. There were 72 patterns, with 4 patterns for each level of similarity per array size.

The data in the Count3 set are the training series from Experiment 3. The number of the dots and similarity (similar to an identical or a different number pattern) were manipulated within subjects. There were 72 patterns, with 6 patterns for each level of similarity at each array size.

Mental arithmetic. The mental arithmetic tasks included a diverse set of problem types. Fits used data broken down by subjects and problem examples.

The data in the Math1 set are from a single-digit multiplication task taken from Experiment 1 of Rickard and Bourne (1996). Either the subjects were shown two digits and asked to calculate the product, or they were shown a digit and a product and asked to divide the product to compute the dividend. RT was recorded as the time between the presentation of the problem and the keystroke of the first digit of the answer. There were 16 problem examples. Problem type (compute product or compute dividend) and range of digits were manipulated within subjects.

The data in the Math2 set are from a three-step arithmetic task (Experiment 1 of Rickard, 1997). Subjects were shown two numbers and asked to calculate their difference, to add 1 to the result, and, then, replacing one of the numbers with the result so far, to compute the sum of it with the remaining original number. RT was recorded as the time between the presentation of the problem and the keystroke of the first digit of the answer. Subjects reported using one of two strategies, recalling the answer from memory or computing the answer (algorithm) on every third trial. We split the data into sets defined by the subject's strategy, using the logistic method described by Rickard (1997),[4] and fit series with more than nine trials.

The data in the Math3 set are from two-digit multiplication and addition tasks, reported as Experiment 1 by Reder and Ritter (1992). Subjects were shown two 2-digit numbers and asked first to indicate which of two strategies, recall (which we label *memory*) or calculate (which we label *algorithm*), they intended to use. They then answered the problem. We fit only to data from the four problems (two addition and two multiplication) that were presented 20 times and excluded one series because it had less than eight correct answers. Also, we fit data for each strategy separately, again excluding the short series.

The data in the Math4 set are from a multistep multiplication task with an initial rapid strategy report, as for the Math3 set (Experiment 2 of Reder & Ritter, 1992). We fit only data from the four problems that were presented 20 times and excluded one series that had less than eight correct answers. The data were also classified by strategy and fit separately, again excluding the short series.

The data in the Math5 set are from a two-digit task combining multiplication and addition, preceded by a rapid strategy report, as for the Math3 set (Experiment 1 of Schunn, Reder, Nhouyvanisvong, Richards, & Stroffolino, 1997). We fit only the three problems presented 28 times and, again, excluded one series with length less than eight. The data were also classified by strategy and fit separately, again excluding the short series.

Alphabetic arithmetic. Subjects were required to verify equations of the form A + 2 = C, or A + 3 = C, true and false equations, respectively. We broke down the data by subjects and by problem example.

The data in the AA1 set are from Experiment 2 of Rickard (1997). Two factors were manipulated within subjects: addend (3, 5, and 7) and trial type (true/false), with four examples of each type. Data for each of the 24 problems were fit separately. Subjects reported strategy as for the Math2 data set. We also analyzed data for the two strategies separately, excluding series with less than 10 responses.

The data in the AA2 set are from an unpublished experiment by Smith and Mewhort (1994). Three factors were manipulated within subjects: addend (2, 3, and 4), arithmetic operator (+, −), and trial type (true/false), with one example of each type. We fit the data from each of the 12 examples separately.

Visual search. In the visual search tasks, subjects were required to indicate whether or not a target appeared in a visual display. In VS1 and VS3, the target was defined by the relative position of two features; in VS2, the target was defined by a conjunction of colors. Stimuli used for targets and distractors were consistently mapped over trials in VS1 and VS2. Targets and distractors were variably mapped in VS3, and a target cue was given before each trial. Fits used data broken down by subjects and by within-subjects factors.

Data in the VS1 set are from Experiment 1 of Heathcote and Mewhort (1993). Two factors were manipulated between subjects: feature type (brightness or color) and display area (small or large). Two factors were manipulated within subjects: display size (two, four, six, or eight objects) and trial type (target/distractor).

The data in the VS2 set are from Experiment 3 of Carrasco, Ponte, Rechea, and Sampedro (1998). Two factors were manipulated within subjects: display size (2, 6, 10, 14, 18, or 22 objects) and trial type (target/distractor).

The data in the VS3 set are from Experiment 3 of Heathcote and Mewhort (1993). Three factors were manipulated within subjects: display size (two, four, six, or eight objects), target type, and trial type (target/distractor).

Motor learning. In the motor learning tasks, subjects were required to press combinations or sequences of keys in response to compatible stimulus displays. All fits were broken down by subjects. Fits for production of combinations of keypresses used series broken down by combination. Fits to sequence production used data broken down by within-subjects factors, but not sequence, since each subject used the same sequence.

The data in the Key1t, Key1c, Key1k, Key2t, Key2c, and Key2k sets were taken fromVerwey (1996). Subjects executed the same nine-key sequence in response to a compatible display. Only the data where all nine key responses were correct were analyzed. On most blocks of trials, the key sequences were divided into segments (called *chunks*) by pauses between stimulus onset. The segment structure was manipulated between subjects: The nine keystrokes were divided into three equal segments (3:3:3) or into two unequal segments (3:6). For the remaining blocks, there was no pause between segments. Otherwise, the next stimulus occurred immediately after the preceding keypress. Subjects were instructed at the beginning of the 2nd day of practice to use the pauses to group their responses temporally.

Trials from Day 1 and from later days and trials from blocks with and without pauses were fit separately. Fits were performed for the time to complete each keypress (Key1k and Key2k), to complete each chunk (Key1c and Key2c), and to complete the total sequence (Key1t and Key2t).

The data in the Key3 set were taken from an unpublished experiment by Brown and Heathcote (1997). Subjects pressed combinations of one to three keys from a set of four keys in response to a compatible visual display (bright rectangle presented on a screen above the response keys). All 14 possible combinations were practiced in random order.

Learning rules from examples. The data were taken from the three experiments reported by J. R. Anderson, Fincham, and Douglass (1997). Subjects studied examples of the form "Skydiving was practiced on Saturday at 5 p.m. and Monday at 4 p.m.," with the underlying rule being that the second practice occurred 2 days later and 1 h earlier. Subjects then indicated the missing parts of similar examples by clicking on one of a set of choices, using a mouse.

The data in the Rule1, Rule2, and Rule3 sets are from Experiments 1, 2, and 3, respectively. All the experiments started with blocks using eight different rules and required the same part (first or second) to be filled in for each rule. Over groups of blocks, examples were introduced that required the other part to be filled in. The manipulation defined the within-subjects factor practiced/unpracticed rule direction. The data for the unpracticed rule direction came from the rules in each group of blocks that had not previously been seen in the unpracticed direction. Experiment 2 introduced a second within-subjects factor: Rules could have either unique or repeated examples. Experiment 3 made the example repetition a between-subjects factor, with either zero, one, or two repeated examples per block.

For fitting, the data were broken down by all within-subjects factors. However, data were *averaged* over rules, and over examples of rules, for groups of four, eight, and eight blocks for Experiments 1, 2, and 3, respectively.

Mental rotation. Subjects were presented with one of four letters (F, G, P, or R) in either their normal or their mirror-image form and rotated by 0°, 30°, 60°, 90°, 120°, or 150°. The subjects' task was to indicate whether the letter was a normal or a mirror-image form. Letter type, letter orientation, and normal versus mirror image were manipulated within-subjects. Subject age (child or adult) was a between-subjects factor.

Data in the MR1c and MR1a sets—child and adult subjects, respectively—were taken from Kail and Park (1990). The data were *averaged* over letter and block (two examples per block); otherwise, we broke down the series by within-subjects factors. Data in the MR2c and MR2a sets were from child and adult subjects, respectively, and came from an unpublished replication of Kail and Park (Ringland & Heathcote, 1998). The latter data were *not averaged*: Letter type was used to break down the series, and we fit individual trial data, rather than block averages.

RESULTS

Note that, where we report a statistic averaged over data sets, the average was calculated weighted by the number of series in each data set (see Table 1 for the number of series in each data set). Table 2 and Figures 1, 2, and 3 give statistics for individual data sets.

The Shape of the Practice Function

Comparing power and exponential functions. Figure 1 reports the percentage of series in each published^5 and unaveraged data set that were better fit by the exponential function than by the power function. Power and exponential functions provided an equally good fit in 2.5% of the series, and these series were excluded from the calculation of the percentages shown in Figure 1. Of the remaining series, the exponential function provided a better fit than the power function in 82.2% of the cases, ranging from a minimum of 64% for the MS2 data set to a maximum of 93% for the Count3 data set. In every case, we could reject, at the 95% confidence level, the null hypothesis that power and exponential functions were equally likely to win.

Table 2 gives the average across all learning series of the proportion of variance (R^2) accounted for by the power and exponential functions. In every data set, the exponential function accounted for more variance than did the power function. Overall, the average R^2 was .498 for the exponential function and .426 for the power function. The exponential fits provided, on average, a 17% improvement, relative to the power function, ranging from 3.7%

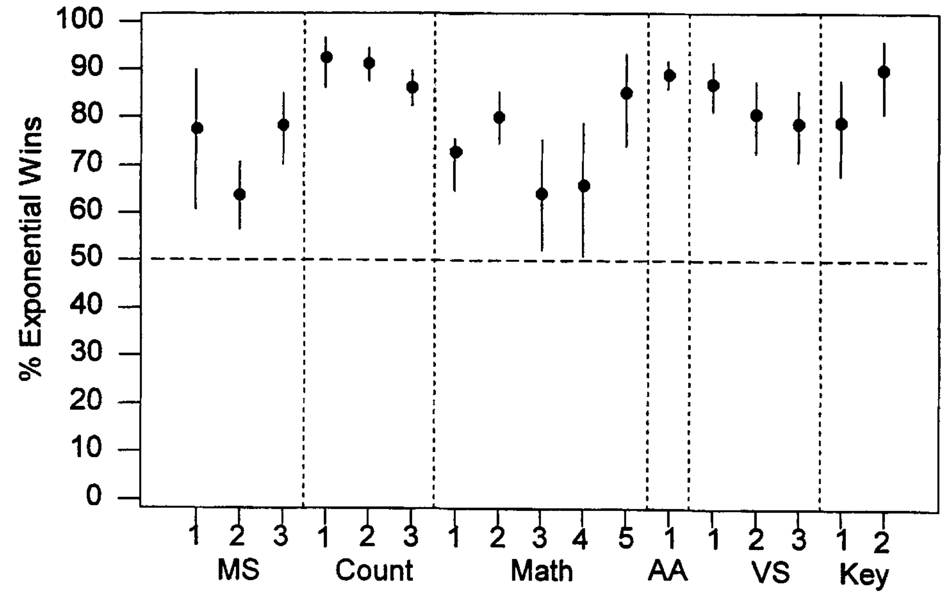

Figure 1. Percentage of cases in which the exponential function provided a better fit than the power function (solid circles), and exact binomial 95% confidence intervals for the published and unaveraged data sets in the survey.

for the Key1t data set to 28.6% for the Count2 data set. In many cases, the absolute increase in R^2 was quite large; that is, the improvement in fit provided by the exponential function was not trivial. Figure 4 reports one such case. Note that the power function's decreasing RLR forces it to approach asymptote very quickly at first and then very slowly, so that it briefly overestimates, then underestimates data in early practice trials. For later practice trials, it overestimates the data. Beyond the range of practice trials measured, it makes a clearly implausible underestimation, that RT reduces to zero.

Sometimes, however, the advantage for the exponential function over the power function was small, and, in most such cases, the R^2 values themselves were also small. Interpretation of small R^2 values for nonlinear regression is tricky, however. The expected value of the nonlinear R^2 can depend on the length of the series. Specifically, the nonlinear R^2 should decrease as series length increases, because extra trials at asymptote add noise but do not add additional signal to the correlation (here, signal means a change in the expected value owing to learning). In the memory search experiments, for example, the practice series were very long, and the absolute values of R^2 were relatively small. Nevertheless, learning, especially early in practice, was quite strong.

Over all the data sets, the difference between R^2 for the power and the exponential functions was highly correlated with their average value ($r = .752$), indicating that the magnitude of the difference was an increasing function of the variance accounted for by learning. Consequently, it appears that smaller advantages for the exponential function are associated with higher levels of noise, rather than being due to a systematic difference between data sets. Taken together with the cautionary note, the correlation shows that the relative increase in R^2 afforded by the exponential function was important in all the data sets.

To assess the strength of learning, we tested each series, using the null hypothesis that the series had a constant mean across practice. Table 2 shows the percentage of series in which learning was significant. For most data sets, the majority of the series show significant learning. Importantly, the two data sets that showed the least significant preference for the exponential (Math 3 and Math4) also showed the weakest learning. Across data sets, the correlation between the percent of exponential winners and the percentage of significant practice effects was $r = .646$. The strong positive correlation again indicates that the weaker advantage for the exponential function shown by some data sets in Figure 1 reflects noise, rather than the form of the underlying learning function.[6]

Comparing APEX and general power functions. Although our analyses, so far, strongly support the exponential function over the power function, the comparison may have been confounded by the effect of preexperimental practice. The power function assumes that preexperimental practice has not occurred, whereas the exponential function does not. When preexperimental practice is important, the power function is forced to estimate too large a decrease in RLR early in practice, and, as a result, the exponential function might dominate. The general power function takes preexperimental practice into account. Hence, to consider a role for preexperimental practice, we compared the fit of the general power function against the fit of the APEX function.

The APEX function can estimate an asymptotic RLR greater than zero, whereas the general power function re-

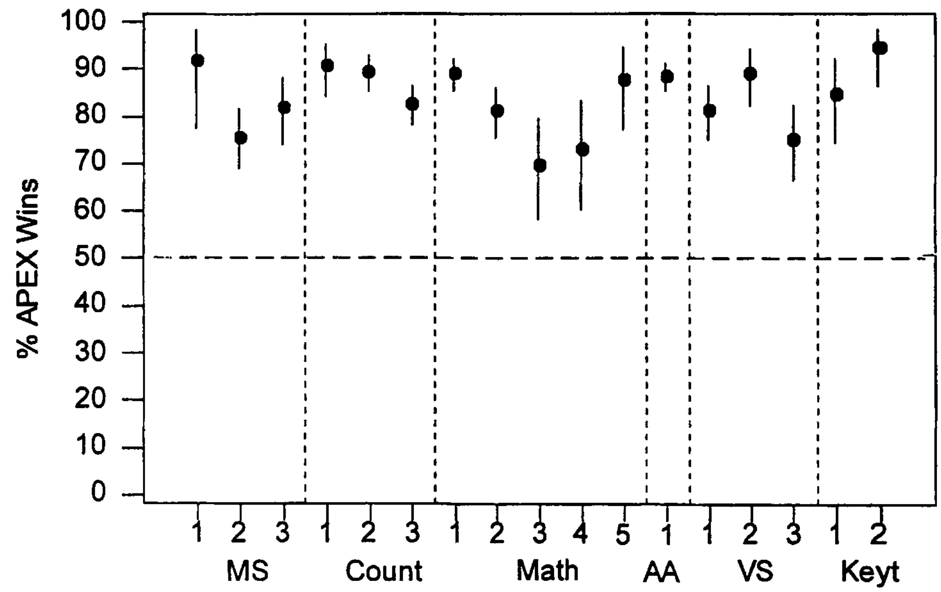

Figure 2. Percentage of cases in which the APEX function provided a better fit than the general power function (solid circles), and exact binomial 95% confidence intervals for the published and unaveraged data sets in the survey.

quires the RLR to decrease to zero across practice. Consequently, better fits for the APEX than for the general power function indicate that the RLR does not decrease to zero (i.e., the function is asymptotically exponential). Conversely, better fits for the general power function indicate that the RLR does decrease to zero (i.e., the function is asymptotically power).

Figure 2 shows the percentage of series that were better fit by the APEX function than by the general power function. In the few cases in which APEX and general power functions provided equally good fits, the general power function was classified as the winner, because these cases corresponded to a power function solution (i.e., the estimate of $\alpha' = 0$ for the APEX function).

Overall, the APEX function won in 84.1% of the series, ranging from a minimum of 70% for Math3 to a maximum of 94% for Key2t. Importantly, all of the cases in which the exponential function was weakest in Figure 1 were more strongly won by the APEX function in Figure 2. Hence, many of the cases won by the power function in Figure 1 (reflecting a decrease in RLR early in practice) do not support a further decrease to an asymptotic value of zero later in practice. These results suggest that the fundamental assumption of the power function—asymptotically negligible RLRs—is not supported by any of the data sets in this survey and that accounting for prior practice is not sufficient to rescue the power law of practice.

Table 2 gives the average R^2 values for APEX and general power functions for each data set. In every data set, the APEX function accounted for more variance than the general power function. Overall, the average R^2 was .507 for the APEX function and .498 for the general power function. The APEX fits provided, on average, only a 2% improvement relative to the general power function. The small difference is to be expected, since the APEX and the general power functions are quite flexible and, hence, able to imitate each other. However, the advantage for the APEX function is very reliable, being evident in both the mean R^2 values and the number of individual learning series winners for every data set. Furthermore, on average, the general power function provided no improvement in fit over the exponential function, despite its extra parameter and ability to mimic exponential data. These results suggest that the better fit of the general power function over the power function is due to its ability to mimic the exponential function, rather than to the effect of preexperimental practice.

Across data sets, the correlation between the percentage of APEX winners and the percentage of significant practice effects was $r = .405$. The value was reduced somewhat by ceiling effects reflecting a very strong preference for the APEX function in some data sets. The difference between R^2 for the general power and the APEX functions was also highly correlated with their average value ($r = .81$). Taken together, the results support the argument that the weaker advantage for the APEX function shown by some data sets in Figure 2 and Table 2 reflects noise, not the form of the underlying learning function.

Assessing asymptotic performance. As is documented in Table 2, the power function tended to predict implausibly fast asymptotic performance. By *implausibly fast*, we mean an estimate of asymptotic expected RT ($\hat{A}_P$ or $\hat{A}_E$) less than 150 msec. Luce (1986) notes that "minimal [simple] visual reaction times are of the order of 180 msec" (p. 63) and that "choice reaction times . . .

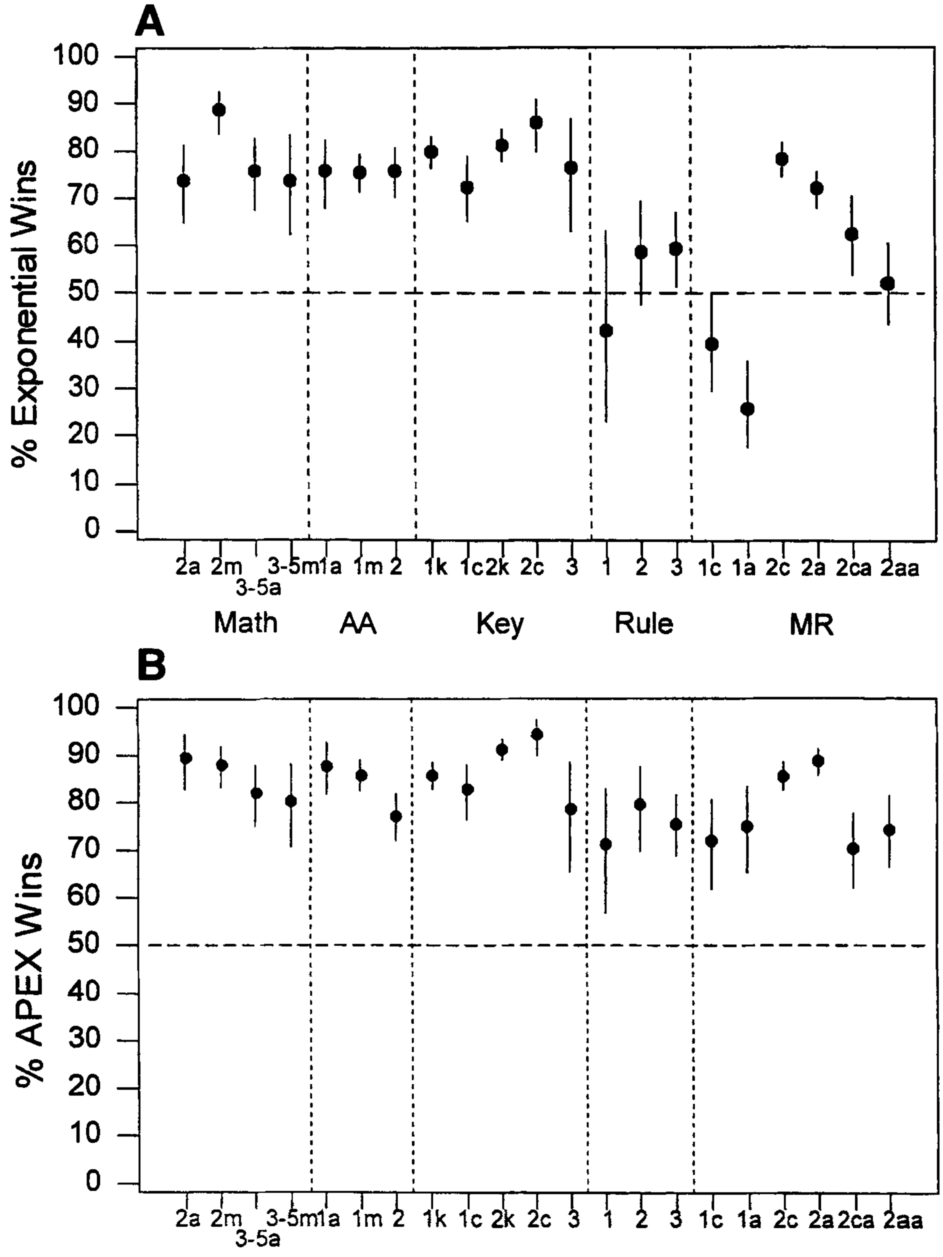

Figure 3. Percentage of cases (solid circles) in which (A) the exponential function provided a better fit than the power function and (B) the APEX function provided a better fit than the general power function, and exact binomial 95% confidence intervals for data sets in the survey not presented in Figures 1 and 4.

are slower than the comparable simple ones by 100 to 150 msec" (p. 208). Hence, a criterion of 150 msec is a conservative estimate of plausible expected RT after extensive practice. Nevertheless, 87.7% of the estimates of asymptotic performance derived from the power function were less than 150 msec, as compared with only 16.1% from the exponential function.

Underprediction of asymptotic performance can be anticipated in a few cases, either because of noise associated with individual subject and condition series or because the series were too short to measure asymptotic performance adequately. The power function's systematic tendency toward underprediction, however, is a symptom of serious misfit. The reason is straightforward: The power function requires a large decrease in RLR from the beginning to the end of practice. Apparently, the power function cannot both match the RLR occurring early in practice and maintain a large enough learning rate late in practice to predict plausible asymptotic performance. Figure 4 illustrates such a case. Note that the power function's slow

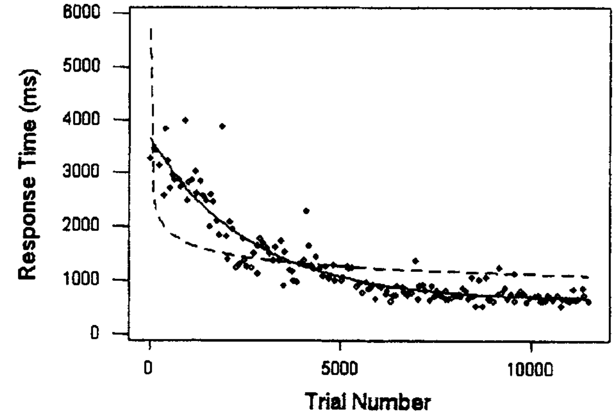

Figure 4. A learning series from the Count3 data set (Subject 2, Stimulus 39, "enemies" trials). Also shown are the best-fitting exponential (solid line; A_E = 840.56, B_E = 3,800, α = .00142, R^2 = .576) and power (dashed line; A_P = 0.00, B_P = 17,037, β = .33922, R^2 = .472) functions.

approach to its (implausibly fast) asymptote means that the underprediction is not evident in the range of practice trials measured.

Allowing for preexperimental practice greatly improves the frequency of plausible asymptotic performance estimates. Overall, the general power function predicted an asymptotic performance of less than 150 msec in only 35.3% of the series. Unlike the power function, the general power function does not have to predict a large decrease in RLR from the beginning to the end of practice, although is does still predict that the RLR eventually decreases to zero. The improvement, however, required large estimates of the effect of preexperimental practice. Overall, the estimates of preexperimental practice were 90.4% of the length of practice series, indicating that almost half of the learning relevant to the experimental tasks occurred prior to the experiment.

Allowing for a constant RLR greater than zero later in practice but a decreasing RLR early in practice increased the frequency of implausible asymptotic performance estimates. Overall, the APEX function predicted an asymptotic performance of less than 150 msec in 27.4% of series. The result raises the possibility that an exponential function is biased to overpredict asymptotic performance, at least when the RLR decreases early in practice. Because the exponential function must predict that the RLR is constant throughout practice, a decrease in the RLR early in practice causes overprediction of the RLR later in practice and, hence, larger asymptote estimates. As was noted earlier, underprediction of the asymptote may occur because of noise and because practice was not carried on sufficiently to obtain an accurate estimate of asymptotic performance. The results for the APEX function suggest that this may have been the case in about a quarter of the series contained in the published data sets.

Three parameters or four? Parsimony suggests that we should prefer a simpler model to a more complex model, if the complex model does not provide a better explanation of the data. As was noted previously, the general power function provides, on average, no improvement in R^2, as compared with the simpler exponential function. Hence, parsimony suggests that we should prefer the exponential function to the general power function as the law of practice.

The simple three-parameter exponential function also fits almost as well as the more flexible four-parameter APEX function. Indeed, in 49% of the data series, the fits of APEX and exponential functions were equal, because the estimate of β' for the APEX function was zero. However, the APEX function did provide, on average, a 2.4% improvement in R^2, relative to the R^2 for the exponential function, with values ranging from 0.6% for the Count1 data set to 7.2% for the Math4 data set.

The latter results indicate that the RLR is either constant throughout practice or decreases slightly early in practice but then remains constant. Over data sets, the difference between R^2 for the exponential and the APEX functions was positively correlated with their average value (r = .334). The correlation suggests that weaker improvements of the APEX function are due to noise, rather than to the underlying shape of the practice function. However, the association is much less than that from comparison of the exponential function with the power function and that from comparison of the APEX function with the general power function. Hence, other factors may be at work. The data sets that show the largest improvements in R^2 for the APEX function (Math2, Math3, and Math4) all come from experiments designed to examine a mixture of algorithmic and memory-based processes. Consequently, a decrease in the RLR early in practice—hence, the need for an APEX law of practice rather than for an exponential law of practice—may often reflect a mixture of strategies.

Individual significance tests. For each series, we compared the fit of the APEX function against that of the power and exponential functions, using change-of-R^2 tests (see Table 2). Overall, goodness of fit decreased significantly when the exponential parameter (α') was fixed at zero in 65.7% of the series, ranging from 6.3% for the Math3 and Math4 data sets to 96.4% for the VS2 data set. By contrast, goodness of fit decreased significantly for only 15% of the series when the power (β') parameter was fixed at zero, ranging from no significant decreases for the Math5 data set to 36.1% for the Key1t data set. The difference in rates supports the previous evidence favoring an exponential function over a power function.

A very strong positive correlation was found between the percentage of significant practice effects and the percentage of significant α' parameters (r = .85). The correlation suggests that cases in which α' was not significant were due to higher noise levels. The correlation between the percentage of significant practice effects and the percentage of significant β' parameters (r = .566) was weaker. The weaker correlation suggests that some cases in which β' was not significant were due to noise but that, in many other cases, β' was not significant because the underlying practice function did not have a power component.

Only 7.8% of the series had both α' and β' estimates significantly greater than zero (i.e., required the full APEX function). The data set that provided the strongest evidence for a power component (Key1t) also provided the strongest evidence for the full APEX function. The percentage of series with significant β' estimates was highly correlated ($r = .879$) with the percentage of series in which both α' and β' estimates were significant; hence, even when the RLR decreased early in practice, the asymptotic RLR did not decrease to zero. The one exception was the Math2 data set; here, 21.4% of the series with a significant β' estimate did not have a significant α' estimate. Even in the Math2 data set, however, significant α' estimates predominated over significant β' estimates.

The individual significance tests support results reported earlier, in suggesting that the exponential function provides a parsimonious model of the law of practice in most cases. Strong evidence for the full APEX function was obtained in only a minority of the series and paradigms. Where evidence existed for a power component, it was usually associated with simultaneous evidence for an exponential component. Hence, a power function alone does not provide a good model of the law of practice, because it wrongly predicts that the RLR decreases to zero with practice.

The Effect of Aggregation

The analyses presented so far used data sets from individual learners and learning conditions to avoid confounding the shape of the function with the effects of aggregation over series with different learning rates. However, some level of aggregation could still have occurred, owing either to a mixture of strategies or to a summation of times for a series of responses. In the following sections, we examine the effects of such aggregation in detail. We then examine the effects of aggregation in data sets that were only available as an average across within-subjects conditions.

Mixed strategies? Rickard (1997) recently suggested that the power law of practice fails because subjects use a mixture of algorithmic and memory-based processes, especially early in practice. Although he maintained that both algorithmic and memory-based processes individually follow a power function, he argued that the mixture does not follow a power function. He also speculated that an advantage for an exponential function over a power function might reflect the mixing of algorithmic and memory-based processing, both of which separately follow a power function (Rickard, May 1997, personal communication).

We tested his suggestions, using data sets from Reder and Ritter (1992), Rickard (1997), and Schunn et al. (1997). In all three studies, subjects reported their processing strategy. For our analysis, the data sets from the latter two papers (Math3, Math4, and Math5) were combined, since each data set by itself had few series.

The results of fits, subdivided by processing strategy, are presented in Figure 3 (in the sections labeled "Math" and "AA"). Both algorithm (75.1%) and memory (79.2%) series were better fit by the exponential than by the power function (77.8%, overall). An advantage for the APEX function over the general power function was even stronger: 86.1% overall, with 86.4% of the algorithm series and 85.9% of the memory series won by the APEX function. As is indicated by the 95% confidence intervals in Figure 3, the preference for exponential and APEX functions was highly significant. We conclude that the better fit for the exponential function is *not* due to a mixture of component power functions.

The results for the series separated by processing strategy are almost identical to the data for the same data sets in Figures 1 and 2, where no distinction was made between processing strategy (83.4% exponential and 84.0% APEX, overall). We take comfort from the similarity, because it argues that mixtures of processing that might have taken place elsewhere in the survey did not distort the tests that we have reported.

Figure 3 also shows the results for a data set from an unpublished alphabet–arithmetic task (AA2) similar to Rickard's (1997) task, except that the subjects were not asked to report their strategy. The results for AA2 also clearly favor the exponential and APEX functions, confirming their dominance in the alphabet–arithmetic paradigm when strategy reports are not required.

Sequential responses. Figure 3 summarizes the effect of a second type of aggregation: summing times from a sequence of responses. In the analyses of Verwey's (1996) data shown earlier, we examined the total time to perform nine sequential keypresses. Because summing across individual keypresses may have distorted the form of the practice function, we looked at the separate responses.

As Figure 3 shows, the time to produce individual chunks and the time to produce individual keypresses were both fit better by the exponential and the APEX functions than by the power and the general power functions, respectively. Combining over data sets from Day 1 and later days, there was a slight reduction in the preference for the exponential for individual keypresses (80.7%) and chunks (79.1%), as compared with the total series (84.7%). There was also a slight reduction in the preference for APEX for individual keypresses (88.6%) and chunks (88.6%), as compared with the total series (89.6%). It is likely that the small reductions reflect greater noise in the keypress and chunks, as compared with the total series. In any case, the results indicate that summing times to perform individual responses did little to distort the form of the practice function in these data.

Averaging across conditions. Seibel (1963) studied subjects who practiced production of all 1,023 possible combinations of 10 keypresses in response to a compatible visual display. Newell and Rosenbloom (1981) reanalyzed the data from one of Seibel's subjects, J.K., from the first 75 blocks of practice. The data were averaged over the 1,023 different combinations of keypresses and blocks. When we reanalyzed the data, using ordinary least squares fitting, the power function ($R^2 = .9858$) provided a better fit than the exponential function ($R^2 = .9584$), but the APEX function ($R^2 = .9902$) provided a better fit than the general power function ($R^2 = .9895$).

Some combinations of keystrokes are harder than others and, as a result, may have been learned at different rates. If so, the shape of the practice function may have been distorted by averaging over conditions with different learning rates. The effect of averaging over conditions with different learning rates is the same as the effect of summing over components of performance with different learning rates; conditions with fast learning rates will dominate the RLR early in practice but will soon approach their asymptote. Hence, late in practice, conditions with slow learning rates will control the RLR. The transition will decrease the RLR of the average function and could yield a better fit for the power function than for the exponential function. However, if learning is exponential for the combinations with slower learning rates, the RLR later in practice will be a constant greater than zero, and, as we found, the APEX function will fit better than the general power function. Further evidence against the general power function in this paradigm is reported by Rosenbloom and Newell (1987b). They found that the exponential function fit better than both the power and the general power functions for a single subject performing Seibel's (1963) task.

Because evidence on the form of the practice function for learning key combinations is both scant and based on relatively short series for each combination, Brown and Heathcote (1997) examined learning of key combinations in a larger sample of subjects with a larger number of trials per combination. Simplifying Seibel's (1963) paradigm, they used only four response keys and the 14 possible keystroke combinations that involved one to three keypresses. By reducing the number of combinations (relative to Seibel's 1,023), they were able to increase the length of the series for each combination.

As is shown in Figure 3, 76.4% of the series were better fit by the exponential than by the power function. The result suggests that the power function fit better than the exponential function for Seibel's (1963) subject J.K. because of averaging. In Brown and Heathcote's (1997) data, 78.6% of the series were better fit by the APEX function than by the general power function. The pattern suggests that the asymptotic RLR did not decrease to zero and is consistent with the results for Seibel's subject J.K. and for Rosenbloom and Newell's (1987b) subject. Hence, the results for key combinations are consistent with the results for other tasks in suggesting that RLR is constant within experimental conditions.

The final analyses presented in Figure 3 concern data sets that were available only as averages across conditions. The first three data sets were from J. R. Anderson et al. (1997); subjects practiced the application of a number of different rules learned from examples. The data were averaged over the different rules and over groups of practice blocks. As is shown in Figure 3A, only one data set produced significantly more exponential than power fits, and one data set displayed a (nonsignificant) preference for the power function.

As is shown in Figure 3B, however, all the data sets were fit significantly better by the APEX function than by the general power function. Overall, the APEX function won in 76.4% of the series from these data sets. The advantage for the APEX function indicates that power components to learning were largely restricted to the early part of practice and that asymptotic learning was exponential. We suspect that the rules differed in difficulty and, hence, in learning rate. Certainly, the pattern reported in Figures 3A and 3B is consistent with the effects of averaging across conditions with different learning rates.

The final averaged data set is from a developmental study of mental rotation of letters performed by Kail and Park (1990). The data were averaged over the four types of letter stimuli and the two presentations of the letters in each practice block. Figure 3 presents the results of analyses separately for children and adults. They show a strong preference for the power function over the exponential function. However, like the other data that included averaging, the APEX function provided a better fit than the general power function, with preference for both children and adults being significant and, on average, 73.4%. As before, the advantage for the APEX function indicates that power components to learning were largely restricted to the early part of practice and that asymptotic learning was exponential.

Because the mental rotation data provide the strongest evidence for the power law of practice of any data set in the survey, Ringland and Heathcote (1998) replicated Kail and Park's (1990) study. As is shown in Figure 3, when analyzed without averaging (MR2c and MR2a), mental rotation, too, significantly favored the exponential function over the power function for both ages. Figure 3 also shows Ringland and Heathcote's data averaged in the same way that Kail and Park's data had been averaged (MR2ca and MR2aa). Although a strong preference for power was not obtained, averaging reduced the number of exponential wins. Moreover, as before, the APEX function provided a better fit significantly more often than the general power function, for both averaged and unaveraged series. In other words, averaging pushes the results toward the power function. We suspect that the remaining differences between Kail and Park's data and Ringland and Heathcote's data reflect differences in learning rates induced by differences in the fonts in which the stimuli were displayed.

Individual Differences

Over almost all of the data sets analyzed, both exponential and APEX functions dominate power and general power functions, but the dominance is not complete. Between 15% and 20% of the learning series were better described by power and general power functions. One explanation is noise. We have provided evidence for this

Table 3
Φ Coefficients Greater Than .2 for All Data Sets in the Survey

Exponential Versus Power				APEX Versus General Power			
Data Set	Factor	Φ	*p*	Data Set	Factor	Φ	*p*
Rule1	Subjects	.664	.638	Rule1	Subjects	.470	.994
MR1	Subjects	.489	.000	Key1t	3:3:3/3:6	.281	1.000
Rule3	Subjects	.440	.861	MR1	Subjects	.280	.449
Math4	Subjects	.434	.674	Rule3	Subjects	.274	1.000
Math1	Subjects	.329	.011	Math3	Subjects	.263	.999
Key1t	3:3:3/3:6*	.307	1.000	Math4	Subjects	.255	.997
MS3	Subjects	.279	.975	MS2	Subjects	.254	.999
VS2	Subjects	.261	.275	VS2	Subjects	.252	.802
Key1c	Subjects	.244	1.000	MS3	Subjects	.235	.997
AA2	Subjects	.231	.881	Key1c	Subjects	.230	1.000
Math5	Subjects	.224	1.000				
Key2t	Subjects	.220	1.000				
Math4	Problem†	.220	.263				

Note—Probability (*p*) refers to a test of the χ^2 value corresponding to each Φ. *Between-subjects factor for chunk size. †Within-subjects factor for problems.

possibility by showing that preference for the exponential and APEX functions increases as noise decreases. Nevertheless, the possibility remains that the power or the general power functions might provide a better description of practice effects in at least some data sets for a minority of conditions or subjects.

We calculated Φ coefficients to quantify the relationship between a preference for exponential and APEX functions and (1) all of the between- and within-subjects factors used in the various tasks in the survey and (2) individual subjects. The Φ coefficient is a nonparametric correlation that ranges between zero and one. It is a linear transformation of the χ^2 statistic used to test the contingency between frequencies in a two-way classification. Larger values of Φ indicate a stronger contingency or association. If the power or the general power function provides a better description of learning for a subset of subjects or a subset of conditions, preference should be systematically related to those subjects and conditions, the corresponding Φ coefficients should be large, and their corresponding *p* values should be small. By contrast, if preference for the power or the general power functions reflects noise, it should not be systematically related to any experimental factor or to particular subjects, and all Φ coefficients should be small. Small values of Φ and correspondingly large *p* values indicate that the null model (i.e., no systematic relationship between factors and preference for a function) provides a good description of the data.

Table 3 presents both estimated Φ coefficients and the significance levels of the associated χ^2 tests.[7] The values of the Φ coefficients were very small in most cases, with a mean value of .11 and a median of only .07. Table 3 reports just the cases, out of more than 200, with Φ estimates greater than .2. All but three of these estimates were associated with subjects (i.e., individual differences), despite the larger number of other types of factors. Two of the three experimental factors with Φ coefficients greater than .2 are between subjects and, so, may be due to individual differences. The only within-subjects factor in Table 3 has a Φ coefficient only slightly greater than .2. Only 2 of over 200 tests indicated significant systematic contingency between function preference and an experimental factor, and in both cases, the experimental factor was a between-subjects factor.

The largest estimates of Φ occurred in the averaged data sets, suggesting that they may reflect individual differences in exponential learning rates across averaged conditions, rather than true power practice effects. The only significant associations were obtained for exponential fits compared with power fits. In the MR1 data set, a single child had exponential wins in all 12 conditions, and two adults had 8/12 exponential wins, whereas the majority of the series were power wins. For the Math1 data set, subjects ranged from all exponential to all power across the 16 within-subjects conditions.

The results suggest that deviation from the general finding of exponential practice is more likely to be associated with individual subjects than with particular within-subjects conditions. The finding reinforces our earlier caution that practice functions from different subjects should not be averaged, not only because variation of learning rates can distort the shape of the average function, but also because an individual subject's learning may occasionally follow a function other than the usual exponential form.

DISCUSSION

Our results can be summarized as follows. The three-parameter exponential function provided a better description of learning than did the three-parameter power function in more than 80% of the cases. The four-parameter APEX function provided a better description than did the four-parameter general power function in about 85% of the cases. Hence, a mixture or a sum of power function processes early in practice cannot explain the power function's loss, since learning was exponential later in practice, when a single process should predominate. In experiments that identified algorithmic and memory-based components, learning in both components was better described by the exponential and the APEX functions than by the power and the general power functions, respectively. The four-parameter general power function provided no improvement over the simpler three-parameter exponential function. Hence, the power function did not lose to the exponential function because of the effects of preexperimental practice.

In about half of the cases considered, the four-parameter APEX function provided no measurable improvement over the three-parameter exponential function, indicating that the extra flexibility implied by a fourth parameter was not needed. Where the fit was improved, the improvement largely reflected a decrease in RLR early in practice. Later in practice, learning was exponential, as is indicated by the dominance of the APEX function over the general power function. Hence, the most parsimonious assumption appears to be that practice produces a simple exponential improvement and a constant RLR, with the caveat that some change in the RLR may occur early in practice.

Despite the ability of nonlinear practice functions to imitate each other, determining the form of the practice function, at least to a first approximation, has not proved to be an insoluble technical issue. For the theoretically important comparison of exponential and power functions, preference for the exponential function was clear and significant in all paradigms. The improvement in fit provided by the exponential function, relative to the power function was, in many cases, quite large, with an average value of 17%. Consequently, the description of practice effects by a power function is often substantially in error (see, e.g., Figure 4).

The estimated parameters of the power function are also misleading because of the problem of imitation. We have shown that at least one parameter of fitted power functions, the asymptote, is usually an underestimation. It is likely that underestimation of the asymptote reflects imitation of a constant RLR, especially later in practice. Because parameter estimates tend to be correlated, it is likely that underestimation of the asymptote distorts estimates of the other parameters of the power function. In short, it is likely that parameter estimates from the power function are unrelated to the psychological processes underlying learning or, at best, related in a complex way determined by the best imitation that a fitting algorithm can find.

The minority of cases in which the exponential function was not the best description appear largely to be due to random variation. Preference for the exponential function was larger in paradigms with higher signal-to-noise ratios, and the frequency of exceptions to the exponential rule was not systematically related to within-subjects conditions. However, a small minority of individual subjects did show systematic deviation from the exponential rule.

Overturning an empirical law requires a high standard of evidence. Our results are clear that the most parsimonious law of practice is the exponential function. Nevertheless, it would be unwise to accept the exponential function without question. Rather, it should form a baseline for comparison with other possible forms, and comparison should be carried out by using individual subjects' data via nested-model tests. Testing should not be carried out on averaged data, since averaging distorts the form of the practice function (R. B. Anderson & Tweney, 1997; Brown & Heathcote, 2000; Myung et al., in press). In the following sections, these techniques will be examined in more detail. We will then consider the implications of an exponential law of practice for theories of skill acquisition.

Measuring Practice Functions

Averaging and practice functions. Averaging usually distorts the form of the practice function. Our analyses for individual subjects and conditions stand in marked contrast to Newell and Rosenbloom's (1981) results for averages over subjects and conditions. Empirically, we found that averaging over within-subjects conditions can produce a bias in favor of the power function. We also found that practice curves for a small minority of subjects may differ from the usual exponential form. Consequently, as Newell (1973) remarked, averaging "conceals, rather than reveals. You get garbage or, even worse, spurious regularity" (p. 295).

Our findings reinforce long-standing analytic results showing that the learning curve for an arithmetic average need not have the same form as the functions contributing to the average (e.g., Estes, 1956; Sidman, 1952). They also confirm that these results are not a mathematical nicety: Arithmetic averages can be biased in favor of a power function and against an exponential function in real data. Myung et al. (in press) have explored the question of why arithmetic averaging of nonlinear functions distorts the average curve and have shown that other averaging techniques are required when dealing with nonlinear models; the appropriate average depends on the nature of the functions to be averaged. Our empirical results and the analytic results converge on the same conclusion: Averaging cannot succeed without first taking into account the form of the functions to be averaged. Researchers can no longer afford to ignore or, worse, to dismiss the effect of averaging as being irrelevant to real data from the paradigms used in the study of learning.

Some variability among component learning rates is necessary for distortion of power and exponential function averages. When component learning rates are exactly equal, the average has the same functional form, and its parameters equal the average of the component's parameters, at least for purely deterministic functions. When the component learning rates vary, neither condition need apply (see Myung et al., in press). The degree of distortion is proportional to the degree of learning rate variability. In particular, for averages over exponential functions, a power-like decrease in the RLR will occur if the sample contains fast and slow learners.

Rickard (1997) suggested that averages do not distort practice functions, because the parameters of learning functions "do not have extremely large variance (a condition that probably holds in most real data sets)" (p. 295). Our experience, both with real data sets and with numerical simulations, differs. Most data sets in the survey contained both fast and slow learners and, sometimes, fast and slow learning conditions, with learning rates often varying over several orders of magnitude. Subjects or conditions that showed a substantial improvement with practice, but at a slow learning rate, were particularly likely to bias the fit toward the power function in the average. Not only were large variations in learning rates regularly observed in the survey data, but also simulations (Brown & Heathcote, 2000) show that the amount of variation in learning rates required to produce significant bias is often as little as one order of magnitude. Hence, we think it is dangerous to assume that the parameters of learning functions are sufficiently homogenous to avoid substantial averaging distortion.

Geometric averaging (averaging on a logarithmic scale) has been widely suggested as a solution to the problem of distortion (e.g., R. B. Anderson & Tweney, 1997; Rickard, 1997). However, geometric averaging preserves the form of averages of power functions and averages of exponential functions only if the asymptote is negligible. It is unreasonable to assume a negligible (near zero) asymptote for practice functions, because performance is limited by physical constraints, such as neural integration time and motor response time. Geometric averaging, therefore, will still distort the form of the average practice function. Although the distortion may be less than that for arithmetic averages (although Brown & Heathcote, 2000, found that the benefits were negligible for the data typically collected in practice paradigms), its effect may differ across experimental conditions or groups of subjects, potentially confounding comparisons.

In general, the scale on which averaging should take place is data dependent. With geometric averaging, the scale depends on the (usually unknown) asymptote parameter, which must be subtracted from the data before logarithmic transformation. Consequently, practice functions must be fit to individual data to estimate the asymptote before averaging can be safely performed. Even then, averaging is problematic, because the asymptote is an expected value and, hence, will be greater than some data points, at least when a sufficiently long practice series is measured to produce a reliable estimate of the asymptote (cf. our critique of Newell & Rosenbloom's, 1981, fitting methods). Subtracting the asymptote estimate from such data points produces a negative number for which the logarithmic transformation is undefined. Hence, even when the appropriate transformation is known or can be estimated, averaging raw data, which is always contaminated by noise, will not be useful when fitting power and exponential functions.

The only case in which averaging is safe is when learning rates vary little between the components of the average. Given that individual analysis is needed to determine when averaging is safe (i.e., to determine the individual learning rates), analysis of the averaged data adds little of value. The survey, however, is encouraging for the study of individual learning. Individual learning effects were often very large, so that averaging was not needed to minimize the effects of trial-to-trial variability. In paradigms with large learning effects, nested-model tests on individual curves were usually decisive. Typically, they indicated the need for an exponential component only, but in a small number of cases, they implicated a power component or both components simultaneously. Such nested-model tests not only provide an inferential basis for conclusions about the form of individual learning curves but also provide a method of identifying unusual cases.

Despite these problems, most researchers would agree that "average data are useful because they often reveal general trends" (J. R. Anderson, personal communication, June 1999). Given its utility, averaging is not likely to surrender its place in data analysis; nor should it. Averaging itself is not intrinsically problematic; biases are only introduced when the scale on which data are averaged does not match either the data or the analyses. Because of noise, averaging of raw practice data is unlikely ever to be useful. However, analysis of average parameter estimates—such as performing an analysis of variance on the parameters of the best-fitting practice functions, as determined by individual analysis—remains viable (e.g., Heathcote & Mewhort, 1995). Such analyses maintain the beneficial effects of averaging, such as reducing noise and revealing general trends, without introducing the systematic distortions produced by averaging raw data. Representations of average performance can be obtained by plotting the practice function with parameters equal to the average of the parameters for individual practice functions.

Nested models and relative learning rates. Detailed consideration of nested-model testing is beyond the scope of the present work (we recommend Bates & Watts, 1988, for further reading). However, a few points of clarification are in order. The benefits of nested-model tests are not limited to practice functions, and useful higher order nesting functions are not limited to the APEX function. Commenting on the form of the forgetting function, for example, Wickens (1998) suggested the use of higher order functions that isolate theoretically important characteristics in separate parameters. The idea is that theories can then be tested by measuring the effect of experimental manipulations on parameter estimates. We chose the APEX function because tests of the α' and β' parameters determined the contributions of theoretically important power and exponential components. Other functions, such as the Weibull, favored by Wickens,[8] or perhaps a sum or mixture of exponential functions may prove useful in other applications. Again, nested-model tests provide an inferential basis for determining a parsimonious form for the function and for testing theories.

An important feature of our work is the use of relative learning to compare practice functions that have different mathematical forms. Wickens (1998) used the hazard rate for forgetting functions in much the same way that we use RLR. RLRs and hazard rates are defined on different measures—expected **RT** and probability of forgetting, respectively—but are otherwise identical.

The importance of the RLR and the hazard rate suggests the desirability of a direct estimate that does not assume a parametric function. However, as Wickens (1998) notes, "Although the use of empirical estimates of the hazard function to select among candidate functions or explanations is attractive, adequate precision is hard to obtain" (p. 382). Nonparametric RLRs can be obtained by dividing estimates of the derivative, such as the difference between RT for adjacent practice trials, by an estimate of the expected value of **RT**. Like other researchers (e.g., Luce, 1986, pp. 60–63), we have found such methods to be inefficient for RT measures. The problem is that such estimates of the derivative tend to be unacceptably noisy, especially with data that are not averaged. Because of these problems, the approach we have taken is fitting simple parametric models with different RLR functions.

Recent advances in nonparametric regression may provide an alternative approach. These techniques reduce the effects of noise by local smoothing with kernels or splines. By estimating the regression function, using local polynomial regression, for example, direct estimates of the derivative are obtainable at every point. These methods have been shown to often provide reliable estimates of differential metrics (Wand & Jones, 1995), such as the RLR.

However, simple parametric functions, particularly the exponential function, provide an important advantage for empirical investigations: They allow the estimation of a single learning rate parameter that applies for all levels of practice. Because the practice function's slope changes with practice, slope cannot provide a single-parameter summary of learning rate. The exponential function's α parameter does provide such a summary by assuming that slope is proportional to the amount left to be learned. The β parameter of the power function also provides a single-parameter summary by assuming that slope is hyperbolically related to the amount left to be learned. However, the results of our survey suggest that the power function does not provide an accurate model of practice effects, and so inferences based on estimates of β may be misleading.

Fitting methods and response time distribution. Least squares fitting, as used in the survey, assumes normally and independently distributed residuals around the expected value function (i.e., the practice function). This assumption is violated by RT data. RT distributions are usually positively skewed, and their means and variances are often positively correlated across conditions (Luce, 1986). Changes in the residual distribution with practice have been analyzed previously (e.g., Logan, 1988; Rickard, 1997), and large decreases in standard deviation with practice have been observed.

These violations of the assumptions of least squares fitting may have biased our results. To check for such bias, we developed a more sophisticated regression technique (APEXL) and applied it to the survey data. Space restrictions do not allow a full explication of the APEXL technique or of the results of its application. Briefly, APEXL fitting uses a special case of Box and Cox's (1964) two-parameter transformation family and implements Carroll and Ruppert's (1988) "transform both sides" approach to regression. It simultaneously estimates both the expected value function's parameters and a data-dependent transformation parameter. The transformation is a shifted logarithm, $\ln(\mathrm{RT} - \lambda)$, where the transformation parameter, λ, is an estimate of the lower bound of RT distribution. APEXL fitting iteratively reweights residuals during fitting, to maximize normal distribution and homogeneity of variance on the transformed scale.

The model underlying APEXL fitting is that RT follows a lognormal distribution. The lognormal distribution is positively skewed and has been found to provide a good account of RT data (Ratcliff & Murdock, 1976). The APEXL model assumes that lognormal distribution is combined multiplicatively with the expected value function, so that RT mean and variance are positively correlated. A detailed analysis by Heathcote and Mewhort (1995) found that the model provided a good description of RT distribution for Heathcote and Mewhort's (1993) visual search practice data. When we applied APEXL fitting to the unaveraged data sets[9] in the survey, 79.0% of the practice series were better fit by exponential than by power functions. Hence, the APEXL analysis supports the conclusion from the least squares analyses that the exponential function is the best simple candidate for the law of practice. We are confident, therefore, that our results were not biased by violations of the assumptions underlying least squares fitting.

Although the results of APEXL and least squares fitting are consistent in their selection of the best practice function, APEXL fitting has a number of advantages over the ordinary least squares approach. APEXL fitting simultaneously estimates the median function along with the expected value function (the two are identical on the transformed scale, since residuals are normally distributed). It provides not only estimates of the expected value function's parameters, but also characteristics of variability around the expected value function, such as the variance and lower bound of the lognormal distribution. When its assumptions hold, APEXL fitting is more efficient than ordinary least squares fitting, a critical advantage when fitting noisy individual data. The assumptions of nested-model tests are better fulfilled on the transformed scale, and so, Type 1 error probability is better estimated. Finally, the form of the expected value function is the same on the transformed and the natural scale, a property that is not true of the more commonly used power transformations (Miller, 1984). Heathcote, Brown, and Mewhort (2000) examine the performance of the APEXL technique in detail, both for data from the survey and for simulated data.

Approaches such as APEXL fitting help to illuminate interesting properties of practice data sets beyond the expected value function, such as the change in RT variance as a function of practice. Such properties can provide useful constraint for theories of skill acquisition (Logan, 1992; Rickard, 1997). Proper measurement of variance, however, relies on prior estimation of the expected value function. Previous attempts to measure variance as a function of practice, for example, used variance estimates calculated from successive blocks of raw data (Kramer et al., 1990; Logan, 1988). Block variance has two components: variance around the expectation function and variance caused by a decrease in expected values across the block. The contribution to block variance from the change in expected value function is unequal in different practice blocks. Early in practice, the expected value function changes rapidly, and, as a result, block variance is greatly inflated by it. After extensive practice, however, the function is relatively flat, so block variance is an almost pure measure of variance around the expected value function.

Predictions from skill acquisition theories have been derived for pure variance around the expected value function (e.g., Logan, 1988, 1992). Consequently, these theoretical predictions cannot be tested by measurement of block variance, which does not purely measure variance around the expected value function. The problem can be remedied by analyzing residuals obtained by subtracting the expected value function from the raw data, but the correct expected value function must be determined first. A similar procedure is required for measurement of other interesting properties of practice data sets, such as autocorrelation between responses. Hence, the results of our survey, which bear on the form of the expectation function, provide the first step in the proper assessment of these other interesting measures.

Before closing this section, we wish to caution against a widely used technique for fitting practice functions: log–log and log–linear plots. In detail, the goodness of fit of power and exponential functions are determined by comparing the linearity of data in log(RT)–log(N) and log(RT)–N plots, respectively. Such fits implicitly assume that the asymptote of the practice function is zero. The assumption is not only wrong but also produces a bias in favor of the power function. The power function approaches its asymptotic value more slowly than the exponential function, and so, its fit is less affected by an underestimated asymptote. In the survey data, small estimates of the asymptote were more commonly associated with the best fits for the power function than with the best fits for the exponential function. Hence, our results show that it is likely that assuming a zero asymptote will hurt the fit of the exponential function more than the fit of the power function. It is also likely that misspecification of the asymptote will make estimates of other practice function parameters, such as learning rates, difficult to interpret, at best, and misleading, at worst, because parameter estimates are usually correlated.

Theoretical Implications

Our results repeal the power law in favor of an exponential law of practice. Many of the data sets included in the survey were collected in order to test specific theories of skill acquisition. We will now consider the implications of an exponential law of practice for these theories. Some of the theories are tied to the power function more tightly than are other theories. Although it is possible for some theories to retain their fundamental assumptions, all require at least some modification in the way these assumptions are applied. The mathematical details of the required modifications are beyond the scope of the present work. In the following, we provide only heuristic details, in order to illuminate the theoretical implications of an exponential law of practice.

Chunking theory. The first response to Newell and Rosenbloom's (1981) power law of practice was their chunking theory, which was later elaborated on by Rosenbloom and Newell (1987a, 1987b). The theory does not exactly predict a power law of practice, but it does predict that RLR decreases to zero with practice. To obtain a decreasing RLR, Newell and Rosenbloom assume (1) that chunks are learned hierarchically, (2) that larger chunks necessarily practice their smaller components every time the larger chunk is practiced, and (3) that no larger span chunk is acquired until all chunks of smaller span are acquired, at least in combinatorial learning environments. A combinatorial learning environment is one in which larger chunks are encountered less often than smaller chunks. The prototypical example is Seibel's (1963) keypress combination task.

Newell and Rosenbloom's (1981) assumptions can be replaced by a single simple assumption to achieve an exponential practice function: Chunks are executed as a single unit and, so, practice only themselves, not their constituents. Newell and Rosenbloom implicitly make a similar assumption by claiming that the execution time for a chunk is independent of its size. Neves and Anderson (1981) also note that such a chunking mechanism, which they call *composition*, produces an exponential practice function.

Even if Newell and Rosenbloom's (1981) theory is not modified, it predicts a decreasing RLR only with practice in a combinatorial learning environment. Arguably, in most of the paradigms examined in the survey, the learning environment was not combinatorial. Hence, chunking theory predicts the observed exponential practice functions. The only data from clearly combinatorial environments come from Siebel's (1963) subject J.K., Rosenbloom and Newell's (1987b) single subject, and Brown and Heathcote's (1997) subjects. Brown and Heathcote's data strongly favored an exponential function, but their paradigm used a smaller set of combinations than did the other experiments; hence, it is likely that the effect of the combinatorial environment was attenuated. It is also likely that Verwey's (1996) paradigm is only weakly combinatorial, since subjects practiced exactly the same sequence on all trials and chunk structure was consistently defined by cues. For simple environments, our results are unambiguous that the RLR does not change much with practice, a pattern predicted by a simple chunking mechanism.

Clearly, more work is required to determine whether complex combinatorial environments yield a decreasing RLR. At present, however, the weight of evidence favors a constant RLR and, hence, the assumption that chunks are seamless units that do not practice their constituents. More work is also required on Newell and Rosenbloom's (1981) derivation of a decreasing RLR (it is not exact). In contradiction of the derivation, several of the simulations of their model presented in Rosenbloom and Newell (1987b) were better fit by an exponential function than by a power function.

Aggregated component theories. Neves and Anderson (1981) suggested a second mechanism based on chunking that produces a decreasing RLR with practice: the summation of many components that learn exponentially. The mechanism is a generic one. For example, Rickard (1997) suggested that response strength for memory-based processing is the sum of strengths of a collection of neural connections that individually learn exponentially, so that the sum is a power function. At a coarser scale, Kirsner and Speelman (1996) argued that many tasks rely on several components, and so, a sum of component learning functions provides a superior model of practice effects.

The critical factor for producing a decreasing RLR in the sum of exponential components is that some components learn more quickly than others. A decreasing RLR occurs for the same reason that averaging across subjects with different learning rates produces a decreasing RLR. Fast learning components produce a large overall RLR early in practice but soon reach asymptote; hence, learning later in practice is controlled by slow learning processes with smaller RLRs.

A modified assumption—that the variation among learning rates is small—is required if theories based on summed exponential components are to predict the exponential practice functions found in the survey. Where all components of the sum have exactly the same learning rate, the sum is exactly exponential. Where learning rates vary, the sum will not be exactly exponential, but if the variation is small, it will approximate an exponential function. Such a model can also accommodate APEX practice functions, if learning rates vary but the smallest learning rates are appreciably greater than zero. Later in practice, the smallest learning rate component or components control learning, so that RLR is asymptotically greater than zero, as in the APEX function. This may account for the evidence for an APEX function that was found in the survey.

It is also possible that a sum of components that learn according to power functions can approximate an exponential or APEX practice function. It is well known that the sum of an infinite number of power functions (i.e., a Taylor series) can approximate nearly any function. However, an infinite number of components is not plausible for psychological processes. For finite sums, approximating an exponential function by the sum of power components also requires assumptions about the weight for each component in the sum. In particular, the weights must vary over at least as many orders of magnitude as there are components. Consequently, power components require justification for the change in weights across many orders of magnitude.

The assumption of power components itself also needs justification. Exponential components can be naturally derived from simple mechanisms. For continuous mechanisms, one need only assume that learning is proportional to the time taken to execute the component. That is, a component that takes longer to execute presents more opportunity for learning. As learning proceeds, the time to execute the component decreases; hence, the learning rate decreases, resulting in exponential learning. For discrete mechanisms, such as chunking, exponential learning occurs for similar reasons. Since responses are produced by larger and larger chunks, fewer opportunities for further composition are available. Similar justifications are needed, if power components are to be plausible.

Rickard's (1997) theory of component power laws also claims that the practice function is the result of an aggregate of component functions. Aggregation occurs through a mixture, rather than by summation, and only two component processes, algorithmic and memory-based processing, are assumed. Our analysis of Rickard's data, combined with analyses of other data sets in which subjects were required to identify responses controlled by algorithmic and memory-based processes, suggests that each component learns exponentially. As has already been discussed, the exponential nature of learning for memory-based processing could result from homogenous learning rates among its component processes.

The mixture assumption in Rickard's (1997) theory can result in a complex change in the RLR with practice, depending on the form of the mixture function. In particular, Rickard's assumption of a logistic mixture usually results in a nonmonotonic change in the RLR of the aggregate, first increasing and then decreasing. This occurs because the change in mixture proportions first accelerates, then decelerates, with practice. Rickard's Math2 data set did provide reasonably strong evidence for an APEX function and, hence, a decrease in RLR early in practice. Explicit fits of the mixture model on the basis of exponential components and consideration of a range of possible mixture functions are required to clarify this issue.

Sums of decaying traces and the forgetting function. J. R. Anderson's (1982) ACT model added a second mechanism, based on strength of learning, to the composition or chunking mechanism already discussed. This mechanism assumes that RT is a linear function of the reciprocal of learning strength. Learning strength equals the sum traces from each practice trial, and the strength of each trace is assumed to decay as a power function of time. The sum of the decaying traces increases approximately as a power function of practice trials, and hence, the model predicts an approximate power decrease in RT with practice. J. R. Anderson, Fincham, and Douglass (1999) develop these ideas more fully and use them to explain slowed performance after a break in practice.

There has been much recent debate on the form of trace decay functions, as measured by forgetting of memorized items at a range of study–test delays. Many of the quantitative issues in this debate reflect the issues we have discussed for practice functions. As we previously noted, Wickens (1998) promoted the use of hazard rates, which are similar to RLRs, to compare and interpret different forms for the forgetting function. His analyses of short-term memory data supported a relatively constant hazard rate and, hence, an exponential function. Most analyses of long-term memory data favor a power function (Wixted & Ebbesen, 1991) or a Weibull function (Rubin & Wenzel, 1996), both of which have decreasing hazard rates. R. B. Anderson and Tweney (1997) suggested that previous analyses of forgetting functions might have been confounded by averaging over subjects. However, Wixted and

Ebbesen (1997) presented a reanalysis of their 1991 data that showed that the power function also applied for individuals.

Recently, Rubin, Hinton, and Wenzel (in press) reported recall and recognition experiments that were specifically designed to determine the form of the forgetting function, using more trials and delays than had been used in previous data sets. Their analysis supported an exponential function for both long- and short-term forgetting, with the rate of short-term forgetting being more than an order of magnitude greater than the rate for long-term forgetting. It is possible, therefore, that decreasing hazard functions for forgetting data are due to a transition between a large hazard rate that is due to a short-term exponential component and a smaller hazard rate that is due to long-term exponential forgetting. Clearly, however, further work is needed to check this speculation.

Given the present uncertainties, the mechanism suggested by ACT to explain the form of the practice function, a sum of decaying memory traces, cannot be ruled out. Indeed, it is an attractive mechanism, because it can potentially unify the results for forgetting and practice functions. However, if ACT is to explain the exponential practice functions found in our survey, and in particular, the constant RLR found later in practice, it must modify its assumptions about the trace decay function.

Instance theories. The theory most closely tied to a power practice function is Logan's (1988, 1992) theory based on the minimum time for race among instance retrievals. As Logan (1995) has acknowledged, "A major goal in developing the theory was to account for the power function speedup" (p. 751). Logan's theory uses a weak learning mechanism, in comparison with other theories of skill acquisition. Each new learning trial speeds performance only because of random variation among retrieval time for traces. Hence, it is suited to predicting the power function's decreasing RLR. Our survey included two of the tasks that Logan's theory was directly developed to account for, alphabet arithmetic and counting patterns of dots. These tasks provided some of the strongest evidence for an exponential function in the survey, with an average improvement in fit of almost 25%, relative to the power function.

Logan (1988) originally claimed that a power practice function could be derived by asymptotic arguments for racing instances with any retrieval time distribution. However, Colonius (1995) showed that the asymptotic argument is flawed, because the asymptotic distribution of minimum times is degenerate, except under linear rescaling that is not justified by Logan's theory. Although Colonius's point reduces the generality of the theory, Logan (1995) countered that a power function is still predicted, using nonasymptotic arguments combined with the extra assumption that instance retrieval times have a Weibull distribution.

To accommodate our results, Logan's (1988, 1992) theory must either assume a different instance retrieval time distribution or add auxiliary mechanisms. One such mechanism could be a race between algorithmic and retrieval processes. However, Logan's (1988) simulations of such a race mechanism did not deviate much from a power function. Second, a race between algorithm and retrieval cannot predict exponential learning later in practice, a clear finding of our survey, unless a substantial proportion of responses continue to be algorithmic throughout practice. Furthermore, we found that both algorithmic and retrieval processes were exponential, in paradigms in which subjects indicated the type of processing that they used. It remains an open question as to whether an alternative retrieval time distribution or other auxiliary mechanisms can allow Logan's theory to predict an exponential practice function.

Logan's (1988, 1992) theory suggests that learning is quite specific and, so, should benefit from preexperimental practice only if that practice is on a very similar task. It is unlikely that preexperimental practice was a strong influence on performance in most of the tasks examined in the survey, because the tasks were probably quite unique in the subject's experience. Logan implicitly assumes a negligible preexperimental practice effect, because he fits the power function, rather than the general power function. Hence, his instance theory cannot take advantage of the improved fit of the general power function. In any case, the general power function achieved no better fit than the simpler exponential function and required average estimates of preexperimental practice almost equal to the amount of experimental practice.

The Evidence Based Rundown Walk (EBRW; Nosofsky & Palmeri, 1997) model is also based on an instance race, but it includes extra mechanisms—specifically, mechanisms reflecting similarity between instances and the accumulation of information via a random walk. The similarity mechanisms make larger estimates of preexperimental practice and, hence, the general power function more plausible than for Logan's (1988, 1992) theory.

The extra mechanisms may also be able to accommodate an exponential practice function. For example, when similarity is negligible, EBRW's predictions follow Logan's (1988, 1992) theory, but the predictions may diverge as similarity increases. Detailed investigation of this issue is beyond the scope of the present work. However, we note that, in Palmeri's (1997) Experiment 2, which used three levels of similarity between stimuli, overall preference for the exponential and APEX functions was high (91.3% and 89.6%) and *decreased* slightly with increasing similarity (96%, 90%, and 88% for exponential preference and 93%, 92%, and 84% for APEX preference). If similarity effects explain exponential practice functions, the power function should be preferred more often with low-similarity stimuli, not with high-similarity stimuli.

Final Word

We do not claim that the practice function is exactly exponential or that theories of skill acquisition must exactly predict an exponential function to be taken seriously. The flexible nature of nonlinear functions means that it is difficult to determine the exact form of the practice function. However, our results indicate that the ability of nonlinear functions to imitate each other does not make the form of the practice function an insoluble technical issue. Furthermore, our results indicate that the form of the practice function can provide a useful constraint for theories of skill acquisition.

The most important characteristic of the exponential function is that it has a constant RLR. Any theory that predicts an approximately constant RLR is supported by the results of our survey. The exponential function defines a baseline against which more subtle theoretical predictions can be tested. It provides a single parameter for the rate of learning that can be used to test hypotheses about factors affecting the efficiency of learning. It also provides plausible estimates of asymptotic performance that can indicate the extent to which learning can improve performance.

If a more flexible practice function is required, our results support the APEX function. The APEX function has the added advantage that it contains the power and exponential functions as special cases, so examination of its parameter estimates and nested-model testing can indicate if either simpler function provides a more parsimonious model. The consistently superior fit provided by the APEX function also suggests that it is likely that nonexponential theories will provide the best fit to data if they predict an RLR that decreases only early in practice, then remains constant at a value greater than zero later in practice.

The difficult nature of discriminating the correct form may have discouraged others from attempting work in this field: Many researchers seem to agree with J. R. Anderson (personal communication, 1999) that "the exact nature of the practice function will never be resolved." However, the results of the present survey are encouraging: They allow a clear discrimination between exponential and power functions as candidates for the law of practice. Clear results were made possible by the willingness of researchers to share their data and by recent advances in the theory of nonlinear regression (Bates & Watts, 1988; Carroll & Ruppert, 1988). Future research using new techniques such as APEXL fitting (Heathcote et al., 2000) and nonparametric regression (Wand & Jones, 1995), coupled with resampling analyses (e.g., Azzalini, Bowman, & Hardle, 1989), promise to take these results further and may allow the form of practice functions to be identified with even greater precision. In order to facilitate this enterprise, we will make practice data used in this survey available on the World-Wide Web.[10] At the time of publication, most survey contributors have agreed to make their data available.

Despite its difficulties, we believe that determination of the mathematical form of empirical laws in psychology is a worthwhile enterprise. Mathematically specified empirical laws both expedite scientific inquiry and guide the development of theory. When we embarked on our survey of the practice function for individual subjects and conditions, we anticipated that the most likely outcome, if the least desirable, would be a variety of function forms for different paradigms and subjects. We were agreeably surprised, therefore, with the consistency of results across the experimental paradigms. The consistency supports Newell and Rosenbloom's (1981) contention that a simple nonlinear function *can* describe practice effects in a broad range of tasks. However, our survey clearly indicates that best candidate for a parsimonious law of practice is the exponential function, rather than the power function.

REFERENCES

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, **89**, 369-406.

Anderson, J. R., Fincham, J. M., & Douglass, S. (1997). The role of examples and rules in the acquisition of a cognitive skill. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 932-945.

Anderson, J. R., Fincham, J. M., & Douglass, S. (1999). Practice and retention: A unifying analysis. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **25**, 1120-1136.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, **2**, 396-408.

Anderson, R. B., & Tweney, R. D. (1997). Artifactual power curves in forgetting. *Memory & Cognition*, **25**, 724-730.

Azzalini, A., Bowman, A. W., & Hardle, W. (1989). On the use of nonparametric regression for model checking. *Biometrika*, **76**, 1-11.

Bates, D. M., & Watts, D. G. (1988). *Nonlinear regression analysis and its applications*. New York: Wiley.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society B*, **26**, 211-246.

Brown, S., & Heathcote, A. (1997). [The law of practice for four choice RT]. Unpublished experiment. Callaghan, Australia: University of Newcastle.

Brown, S., & Heathcote, A. (2000). *Fitting nonlinear models to averaged data*. Manuscript submitted for publication.

Carrasco, M., Ponte, D., Rechea, C., & Sampedro, M. J. (1998). "Transient structures": The effects of practice and distractor grouping on within-dimension conjunction searches. *Perception & Psychophysics*, **60**, 1243-1258.

Carroll, R. J., & Ruppert, D. (1988). *Transformation and weighting in regression*. New York: Chapman and Hall.

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

Colonius, H. (1995). The instance theory of automaticity: Why the Weibull? *Psychological Review*, **102**, 744-750.

Delaney, P. F., Reder, L. M., Staszewski, J. J., & Ritter, F. E. (1998). The strategy-specific nature of improvement: The power law applies by strategy within task. *Psychological Science*, **9**, 1-7.

Estes, K. W. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, **53**, 134-140.

Heathcote, A. (1990). *Learned pop-out in search for relative position*. Unpublished doctoral dissertation, Queen's University, Kingston, Ontario, Canada.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). *Nonlinear regression for response time data*. Manuscript in preparation.

Heathcote, A., & Mewhort, D. J. K. (1993). Selection and representation of relative position. *Journal of Experimental Psychology: Human Perception & Performance*, **19**, 448-515.

Heathcote, A., & Mewhort, D. J. K. (1995, November). *The law of practice*. Poster presented at the 36th Annual Meeting of the Psychonomic Society, Los Angeles.

Josephs, R. A., Silvera, D. H., & Giesler, R. B. (1996). The learning curve as a metacognitive tool. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **22**, 510-524.

Kail, R., & Park, Y.-S. (1990). Impact of practice on speed of mental rotation. *Journal of Experimental Child Psychology*, **49**, 227-244.

Kirsner, K., & Speelman, C. (1996). Skill acquisition and repetition priming: One principle, many processes? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **22**, 563-575.

Kling, J. W. (1971). Learning: An introductory survey. In J. W. Kling & L. A. Riggs (Eds.), *Woodworth and Schlosberg's Experimental psychology* (pp. 551-613). New York: Holt, Rinehart & Winston.

Kramer, A. F., Strayer, D. L., & Buckley, J. (1990). Development and transfer of automatic processing. *Journal of Experimental Psychology: Human Perception & Performance*, **16**, 505-522.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, **95**, 492-527.

Logan, G. D. (1992). Shapes of reaction-time distributions and shapes of learning curves: A test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 883-914.

Logan, G. D. (1995). The Weibull distribution, the power law, and the instance theory of automaticity. *Psychological Review*, **102**, 751-756.

Luce, R. D. (1986). *Response times*. New York: Oxford University Press.

MacKay, D. (1982). The problems of flexibility, fluency, and speed–accuracy trade-off in skilled behavior. *Psychological Review*, **89**, 481-506.

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

Miller, D. M. (1984). Reducing transformation bias in curve fitting. *American Statistician*, **38**, 124-126.

Myung, I. J., Kim, C., & Pitt, M. A. (in press). Toward an explanation of the power law artifact: Insights from a response surface analysis. *Memory & Cognition*.

Neves, D. M., & Anderson, J. R. (1981). Knowledge compilation: Mechanisms for the automatization of cognitive skills. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-55). Hillsdale, NJ: Erlbaum.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information processing* (pp. 283-308). New York: Academic Press.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-55). Hillsdale, NJ: Erlbaum.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random-walk model of speeded classification. *Psychological Review*, **104**, 266-300.

Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 324-354.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1986). *Numerical recipes: The art of scientific computing*. New York: Cambridge University Press.

Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review*, **83**, 190-214.

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 435-451.

Rickard, T. C. (1997). Bending the power law: A CMPL theory of strategy shifts and the automatization of cognitive skills. *Journal of Experimental Psychology: General*, **126**, 288-311.

Rickard, T. C., & Bourne, L. E. (1996). Some tests of an identical elements model of basic arithmetic skills. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **22**, 1281-1295.

Ringland, C., & Heathcote, A. (1998). [The effect of practice on speed of mental rotation: A developmental comparison].Unpublished experiment. Callaghan, Australia: University of Newcastle.

Rosenbloom, P. S., & Newell, A. (1987a). An integrated computational model of stimulus–response compatibility and practice. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 21, pp. 1-52). New York: Academic Press.

Rosenbloom, P. S., & Newell, A. (1987b). Learning by chunking: A production system model of practice. In D. Klahr, P. Langley, & R. Neches (Eds.), *Production system models of learning and development* (pp. 221-286). Cambridge, MA: MIT Press.

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

Rubin, D. C., Hinton, S., & Wenzel, A. E. (in press). The precise time course of retention. *Journal of Experimental Psychology: Learning, Memory, & Cognition*.

Schunn, C. D., Reder, L. M., Nhouyvanisvong, A., Richards, D. R., & Stroffolino, P. J. (1997). To calculate or not calculate. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **12**, 1-27.

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

Sidman, M. (1952). A note on functional relations obtained from group data. *Psychological Bulletin*, **49**, 263-269.

Smith, D. G., & Mewhort, D. J. K. (1994). [*Why practice makes perfect: An analysis of the instance theory of automaticity*]. Paper read at the meeting of the Canadian Society for Brain & Behavior, University of British Columbia.

Strayer, D. L., & Kramer, A. F. (1994a). Aging and skill acquisition. *Psychology & Aging*, **9**, 589-605.

Strayer, D. L., & Kramer, A. F. (1994b). Strategies and automaticity I. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 318-341.

Strayer, D. L., & Kramer, A. F. (1994c). Strategies and automaticity II. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 342-365.

Thorndike, E. L. (1913). *Educational psychology: The psychology of learning* (Vol. 2). New York: Teachers College Press.

Verwey, W. B. (1996). Buffer loading and chunking in sequential keypressing. *Journal of Experimental Psychology: Human Perception & Performance*, **22**, 544-562.

Wand, M. P., & Jones, M. C. (1995). *Kernel smoothing*. London: Chapman & Hall.

Wickelgren, W. A. (1975). Memory storage dynamics. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes* (Vol. 4, pp. 321-361). Hillsdale, NJ: Erlbaum.

Wickens, T. D. (1998). On the form of the retention function: Comment on Rubin and Wenzel (1996): A quantitative description of retention. *Psychological Review*, **105**, 379-386.

Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science*, **2**, 409-415.

Wixted, J. T., & Ebbesen, E. B. (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory & Cognition*, **25**, 731-739.

NOTES

1. For games won, R^2 was .339 for the exponential function and only .253 for the power function. For games lost, R^2 was .183 and .173 for the exponential and power functions, respectively. The fit of the general power function was somewhat better (R^2 of .334 and .185), but it was still less than the exponential function for games won and less than the APEX function (with R^2 of .339 and .186, respectively) in both cases. We especially thank Paul Rosenbloom for sending us these data sets.

2. Analysis of RLRs also shows that Mazur and Hastie's (1978) results, which Newell and Rosenbloom (1981) claim agree with their results in "rejecting exponentials" (p. 34), are not relevant for RT. Mazur and Hastie fit power and exponential functions to rate of response data. A nonlinear transform, the inverse, is required to convert rate to RT. The transform changes the RLR of the exponential rate function to an RLR on the RT scale [$K/(e^{kN} - 1)$, where K and k are parameters greater than zero] that decreases more quickly than the power function's RLR on the RT scale. Hence, Mazur and Hastie's comparison does not test a true exponential function on the RT scale.

3. A related function was proposed by Wickelgren (1975, p. 326) to model retention of memories. Note that a prior-practice parameter analogous to that in the general power function could also be added to the APEX function (i.e., substitute [$N + E$] for N). This five-parameter function nests both the APEX and the general power functions. It was

not used because the general power function was over-parameterized for the practice data sets examined and often provided ill-conditioned objective functions that made minimization difficult. Hence, the five-parameter version of the APEX function could only be more problematic.

4. We also used a number of simpler schemes for partitioning trials with the similar results.

5. The single exception is the unpublished two-alternative forced-choice experiment included in the MS2 data set. These data were included, since they are very similar to all others in MS2.

6. In the limit of no learning, power and exponential functions will both win about 50% of the series each. Hence, a 50% result means either that the true shape of the practice function falls between the shapes of the power and exponential functions or that learning is very weak.

7. The significance of each Φ coefficient was tested with the corresponding χ^2 test. The significance levels for these tests must be treated with some caution, especially for the subjects factor, since they assume independence.

8. The Weibull function is a power transformation of the exponential function. With an exponent of one, it equals an exponential function and, so, has a constant RLR. Exponents greater than or less than one produce increasing and decreasing RLRs, respectively. Wickens (1998) also used the Parto II function. The Parto II function is a special case of the general power function. Our evidence suggests that the general power function is inferior to the APEX function and is plagued by ill-conditioned fitting for practice curves.

9. The averaged data sets were not analyzed with the APEXL technique, because the central limit theorem implies that averages tend to be distributed normally.

10. From http://psychology.newcastle.edu.au/ follow the links to the first author's home page.

(Manuscript received February 1, 1999;
revision accepted for publication June 15, 1999.)