Behavior Research Methods, Instruments, & Computers
2003, 35 (1), 11-21

Averaging learning curves across and within participants

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

We examine recent concerns that averaged learning curves can present a distorted picture of individual learning. Analyses of practice curve data from a range of paradigms demonstrate that such concerns are well founded for fits of power and exponential functions when the arithmetic average is computed over participants. We also demonstrate that geometric averaging over participants does not, in general, avoid distortion. By contrast, we show that block averages of individual curves and similar smoothing techniques cause little or no distortion of functional form, while still providing the noise reduction benefits that motivate the use of averages. Our analyses are concerned mainly with the effects of averaging on the fit of exponential and power functions, but we also define general conditions that must be met by any set of functions to avoid distortion from averaging.

Averaging is perhaps the most common statistical analysis technique used in psychological research today. Researchers average data in order to identify systematic relationships between noisy variables. When the relationships being examined are ordinal, the average is representative of its components. However, modern psychological research has moved beyond the simple determination of orders. Instead, competing quantitative models of individual behavior are tested by comparing their fits to data. Often, the models are complex and nonlinear. One of our aims here is to demonstrate that such models should *not* be compared on the basis of their fits to averaged data; under quite general conditions, the average function has a different mathematical form than do its component functions. Even when the form of the component functions is preserved, the parameters of the average function may not equal the average of the parameters of the component functions.

As early as 1892, Boas questioned the representativeness of averaged growth curves, and like concerns have been reiterated for a variety of learning curves (see, e.g., Bahrick, Fitts, & Briggs, 1957; Bakan, 1954; Estes, 1956, 2002; Kling, 1971; Sidman, 1952; Underwood, 1949). For example, Sidman showed that the average of a finite number of exponential functions could never be exactly exponential in form itself, if the rate parameters of the components differ. Such results for closed-form functions are usually simple to prove—for instance, a similar result holds for the average of power functions.

In the main, empirical analyses in psychology have ignored these problems, perhaps because they do not arise when only ordinal relationships are of interest or when component functions are linear. The potentially misleading effects of averaging may also have been ignored because of the lack of a clear and convincing demonstration that averaging can cause serious errors in conclusions about real psychological data. After all, averaging distortions occurring in real data may be neither substantial nor theoretically misleading. In particular, results such as Sidman's (1952)—that an average of exponential functions could not be exactly exponential—does not rule out the possibility that the average might be close enough to exponential for practical purposes.

Two surveys of skill acquisition data (Heathcote, Brown, & Mewhort, 2000; A. Newell & Rosenbloom, 1981) have provided evidence that distortions due to averaging have led to misguided theory development in psychology. In this paper, we endeavor to augment the averaging literature by providing several results dealing with averages across different curves: (1) a simplified version of an established proof, outlining exactly when arithmetic averages of any nonlinear function will be distortion free, (2) Monte Carlo simulations showing that averages of exponential functions can easily mislead model discriminations, and (3) reanalyses of published data sets, demonstrating that averaging across curves can have strong effects in real psychological data. Historically, the dangers of averaging have often been ignored, probably because the benefits of averaging are so enticing. With this in mind, we also provide some more positive news for the uses of averaging, by showing (4) that there are methods of averaging within curves that can result in little or no distortion but still provide strong noise reduction benefits, and (5) the effect of these methods on published data sets.

Our conclusions are directed most generally at any attempt to average nonlinear functions. When the need arises

Correspondence concerning this article should be addressed to S. Brown, Department of Cognitive Sciences, University of California, SSPA 3181, Irvine, CA 92697 (e-mail: scottb@uci.edu).

to choose specific nonlinear functions for examples, we use exponential and power functions. We chose these two functions because they are simple and well understood and were naturally suggested by the results of A. Newell and Rosenbloom's (1981) and Heathcote et al.'s (2000) studies. We focus on the most common form of averaging, arithmetic, but also address geometric averaging.

AVERAGING ACROSS FUNCTIONS

Initially, we will restrict our analyses to perhaps the most common form of averaging in empirical psychology: that performed across different functions. Averages across participants, across different experimental conditions, or across different items or stimuli all fall within this class. Thomas and Ross (1980) have defined two properties that make averages representative of their components: *functional isomorphism* and *parameter averaging*. An average is functionally isomorphic only when it has the same mathematical form as its components. Functional isomorphism makes sense only if all component functions have the same form (e.g., they all belong to the same family of equations). The second property, parameter averaging, extends functional isomorphism by requiring that the parameters of the average function equal the average of the parameters of the component functions. Parameter averaging makes sense only when functional isomorphism holds. Where both properties hold, we will describe the average function as *representative*.

A necessary and sufficient condition for an arithmetic average function to be representative is that its component functions, which may have an arbitrary nonlinear form, are linear in parameters that vary across components (see Appendix A). All such linear functions can be expressed as a matrix multiplication function (assuming one can identify a suitable basis set for the response space). For the most common case, y: $R^k \rightarrow R$, this means that if y is to have a representative average, it can be expressed as the inner product of a fixed parameter vector $\mathbf{A}$ with a k-vector-valued function of x, $f(x,\theta)$, where θ is a fixed vector of nonlinear parameters. So, any function for which the arithmetic average will be representative may be expressed in the form

$$y(x) = a_1 f_1(x,\theta) + a_2 f_2(x,\theta) + \ldots + a_k f_k(x,\theta).$$

Modern psychology is often concerned with nonlinear functions, and so, according to the result above, averages will *not* be representative, at least whenever their nonlinear parameters (θ) vary. Exponential functions are an important example: Whenever their rate parameters vary across different component curves, the average of those curves will not be exponential in form. However, it may still be the case that the average function is close enough to an exponential form for empirical use. Whether or not an average across exponential functions is noticeably different from exponential form is a question that has received mixed answers in the psychological literature. In a reply to concerns about averaging raised by Anderson and Tweney (1997), Wixted and Ebbesen (1997) reported little averaging distortion in the data of Wixted and Ebbesen (1991), since *both* individual and average retention functions from were better fit by power than by exponential functions. For averaged practice curves, A. Newell and Rosenbloom (1981) found that the power function provided a better fit than the exponential function in data from many paradigms. Heathcote et al. (2000), in contrast, analyzed unaveraged practice curves and concluded that the exponential function provided a better fit than the power function in every case.

Analyses of Simulated Data

In this section, we report the results of Monte Carlo simulations, the aim of which was to reconcile the apparently contradictory results about the effects of averaging reported in the literature. Anderson and Tweney (1997) used simulations similar to ours, but with important limitations. We will define power and exponential functions as in Equations 1 and 2, respectively, with y the criterion and x the predictor. The subscripts P and E are used to differentiate the linear parameters for power and exponential functions, respectively. Where we drop the subscript, we mean the parameter to stand for either power or exponential function parameters, unless otherwise stated. Both functions have one nonlinear parameter, s for the power and r for the exponential:

$$y_P = a_P + b_P x^{-s}, \tag{1}$$

and

$$y_E = a_E + b_E e^{-rx}. \tag{2}$$

Anderson and Tweney (1997) examined only two-parameter versions of the power and exponential functions, with a constrained to be zero in both cases. In the two-parameter case, geometric averaging[1] provides an easy solution to averaging distortion—but one that cannot be extended to other functions in which the asymptote (a) parameters vary. Our simulations extend Anderson and Tweney's work to the three-parameter case, using a variety of curve lengths and numbers of component curves, as well as varying linear parameters among components. These extensions make our simulations a more accurate reflection of the situation in psychological research. In all further descriptions, the exponential functions to be averaged (the *component functions*) will be specified as $y_E = a_i + b_i e^{-r_i x}$, where $x = 1, 2, \ldots, M$ indexes the values of the covariate (e.g., experimental trials), and $i = 1, 2, \ldots, P$ indexes the set of component curves to be averaged (e.g., participants). For a description of the parameters and the numerical methods used, see Appendix B.

Simulation Methods

Simulation 1 averaged two (noiseless) exponential functions, both with $a_i = 0$, and determined whether the average curve was better described by an exponential or

a power function. The spread of the component functions' r parameters was manipulated until the point was found at which the average changed from being more like an exponential function to being more like a power function. Algorithmically, the largest of the two r values, r_{max}, was fixed, and a line search over the other r parameter, r_{min}, was used to locate the value at which the sum-squared deviations from the power function were equal to those from the exponential. We will call the value of the ratio r_{max}:r_{min} at the point at which the average curve changed from more like an exponential function to more like a power function the *distortion ratio*. The distortion ratio provides a measure of the amount of difference in the rate parameters required for averaging to cause sufficient distortion to mislead inferences about functional form based on goodness of fit. These simulations were repeated using many different combinations of r_{max} and of M, although only two component functions were used in Simulation 1 ($P = 2$). The b_i parameters were constrained to be constant across component functions.

Simulation 2 extended Simulation 1 by varying P between 2 and 20, mimicking the process of averaging across larger groups of participants or experimental conditions. The value of r_{max} was again fixed, and a search was performed over r_{min}. The values of the other r_i were logarithmically[2] spaced between r_{max} and r_{min}. These simulations had essentially the same outcome as Simulation 1 (i.e., $P = 2$). In fact, the level of distortion for a given r_{max}:r_{min} ratio with $P > 2$ was always either the same as that for $P = 2$ or even greater than that for $P = 2$.

Simulation 3 extended Simulation 2 by using component functions with values of $a_i > 0$ and values of b_i that varied across component functions (i.e., three parameter functions). These simulations were designed to check whether using constant (and zero) asymptotes and constant b_i values had restricted the results of the previous simulations. Analysis was complicated by the nonzero a_i parameters, which necessitated the use of a nonlinear regression algorithm to fit the power and exponential functions at each iteration of the line search.

The results of the simulations with $a_i > 0$ and varying b_i yielded the same outcome as the simulations with $a_i = 0$ and b_i fixed. This result concurs with Anderson and Tweney (1997), who found that varying the b_i values had very little effect on averaging artifacts (they did not vary a_i). The agreement of results from Simulations 1 and 2 with the results from Simulation 3 also suggests that the use of the two different fitting algorithms (transformation and linear regression for the two-parameter functions vs. nonlinear regression for the three-parameter functions) did not produce different results. This agreement is not trivial, since the two methods assume different error models: additive normal errors for the nonlinear regression and log-normal errors for the transformed linear regressions.

Simulation Results

Although the effect of manipulating component r values did not vary much across the three sets of simulations, it was more complex than might be expected. A typical graph of the distortion ratio (i.e., the value of r_{max}:r_{min} at the point at which distortion became strong enough to cause misidentification of the average curve) is shown in Figure 1 as a function of r_{max} (solid line). There are several important things to note from this graph. First, the value of the distortion ratio initially decreases as the value of r_{max} decreases, indicating that less variability in r parameters is required to cause significant averaging distortion as the maximum rate parameter decreases. Second, the distortion ratio achieves a minimum of approximately 10. That is, averaging can cause strong distortions when the difference between the maximum and the minimum r parameters is only one order of magnitude.

Perhaps the most surprising effect observed in the simulation results is the upturn in the low-r_{max} section of Figure 1; this upturn occurred for every combination of parameters examined. The distortion ratio initially decreases with decreasing r_{max} but then reaches a minimum, after which it rapidly increases to effectively infinite values. This behavior means that, for approximately $r_{max}M < 10$, averaging exponential functions can *never* result in a curve that is better fit by a power function than by an exponential function. A simple graphical explanation of the no-distortion region of Figure 1 is that, when $r_{max}M$ is quite small (and thus, $r_{min}M$ is even smaller), all of the component functions are nearly constant. Averaging any number of (nearly) flat lines with an exponential function will not greatly distort curve form.

Reanalyses of Published Data Sets

To examine the impact of averaging on real psychological data, 17 data sets analyzed by Heathcote et al. (2000) were reanalyzed after arithmetic averaging across participants. Figure 2 displays the results listed in order of the percentage of individual (unaveraged) curves better fit by the exponential. The details of the methods and the definitions of the labels used for each data set are given in Appendix C. There were many experimental

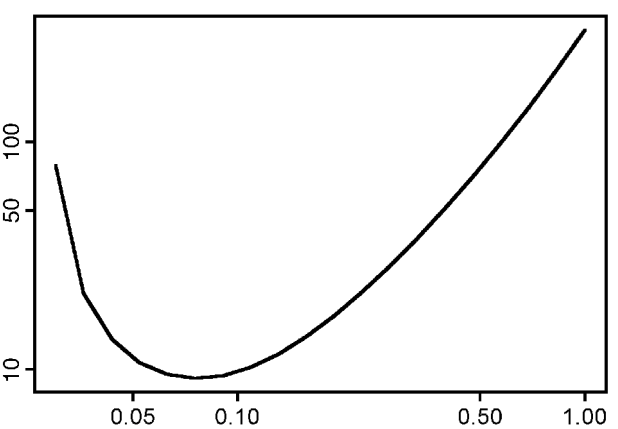

Figure 1. Log-log plot of distortion ratio versus r_{max} for a typical function set (two functions, length 100, $a_1 = a_2 = 0$, $b_1 = b_2 = 5{,}000$).

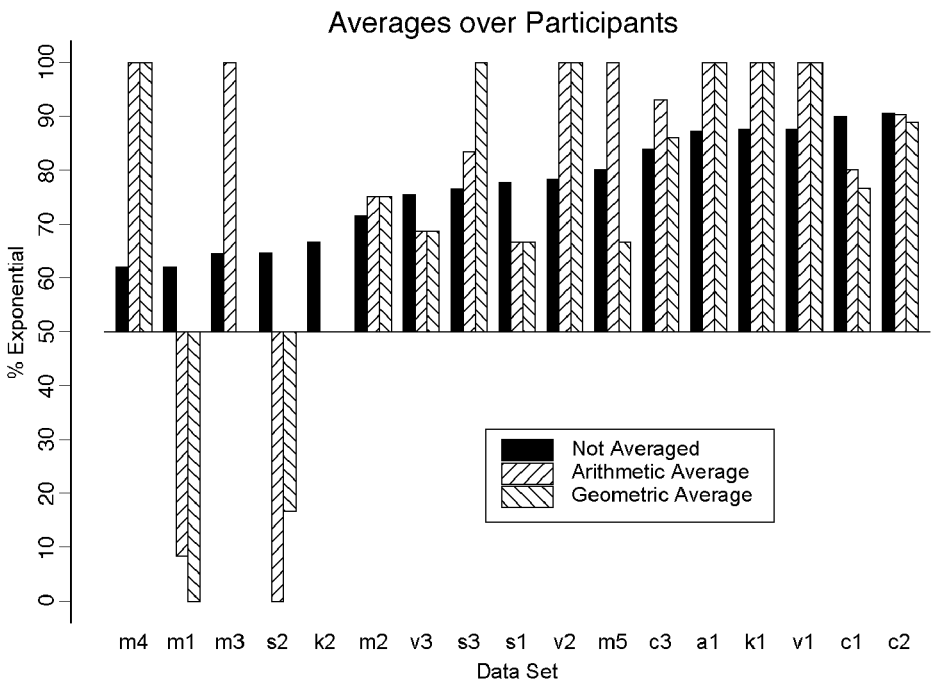

Figure 2. Percentages, over conditions and participants, of fits where a three-parameter exponential function provided a better fit than did a three-parameter power function in unaveraged data, and corresponding percentages, over conditions only, for fits to data arithmetically and geometrically averaged over participants. Note that where bars appear to be absent, this indicates equal (50%) preference.

conditions (between 8 and 260) in each data set, and data were not averaged over these—only across participants within conditions. The results show that averaging over participants has strong and unpredictable effects on model selection. Prior to averaging, all the data sets showed a consistent preference for the exponential function (in 62%–91% of cases, the exponential provided a better fit). In 7 of the 17 data sets, averaging over participants decreased the preference for the exponential; two cases resulted in a strong preference for the power function.

For some data sets shown in Figure 2, averaging over participants increased, rather than decreased, preference for the exponential function. An explanation for this effect is that, as noise increases and dominates the underlying learning curve, both power and exponential functions will provide the best fit equally often. Although this assertion seems intuitively plausible, Myung, Kim, and Pitt (2000) reported a very strong bias whereby the power function provided a better fit to purely random data than did the exponential function in 99% of cases. Myung et al.'s simulations were limited to only single-parameter power and exponential functions (i.e. where a and b were both fixed). Brown and Heathcote (in press) replicated Myung et al.'s results, but when they extended the analysis to exponential and power functions with two parameters (only a fixed) and three parameters varying, they found equal preference in fits to random data. Hence, in Figure 2, which reports the fit of three-parameter power and exponential functions, averaging over participants may have increased the preference for the exponential function because it reduced noise and, so, moved the results away from equal preference.

The conclusion from these analyses is clear: Arithmetic averaging yields unpredictable results, which may confound model comparison. Importantly, there seem to be two competing forces at work: a distorting effect on functional form and, also, a clarifying effect based on noise reduction. When analyzing only averaged data, it is difficult to know which of these two has dominated, and so conclusions about the nature of the unaveraged data are difficult to make. Where issues of individual curve form and, more generally, the fit of any nonlinear model are at stake, data should not be averaged across functions.

Geometric Averaging

Geometric averaging preserves the functional form of power and exponential data, but only when component curves have zero asymptotes. When the asymptote values are large relative to the range of the data, the logarithmic transformation is close to linear across that range, and so geometric averaging is similar to arithmetic averaging and, hence, ineffective. Subtracting asymptote estimates before geometric averaging is often not a practical solution to this problem, not only because the asymptote must be estimated, but also because zero

and negative observations may result from the subtraction, owing to noise,[3] in which case a logarithm is not defined.

However, even when nonzero asymptotes are not corrected, it is possible that geometric averaging may result in averages that are close enough to representative for practical purposes. We performed simulations to investigate this possibility by determining exactly how different from zero asymptotes have to be before geometric averaging is ineffective. When $a_i = 0$, geometric averaging completely removed the bias toward the power function, as was expected. As a_i increased, this benefit decreased. These effects are illustrated in Figure 3 as a function of the ratio a_i/b_i. The ordinate in Figure 3 represents the difference in residual sum squares (RSS) between the best-fitting power and exponential functions for arithmetic and geometric averages of exponential functions. Large negative values indicate no distortion (the exponential fit has small RSS, the power fit has large RSS). The origin of the ordinate axis represents the point at which model discrimination decisions would be reversed.

Arithmetic averaging, for all values of a_i/b_i, causes distortion toward the power function. For values of a_i/b_i less than about 0.3, geometric averaging was relatively effective in removing bias toward the power function. At all values of a_i/b_i greater than this, the bias toward the power function was strong enough to reverse model discrimination choices made on the basis of RSS. For values of a_i/b_i greater than approximately 0.5, the bias due to geometric averaging was almost as large as that due to arithmetic averaging. This result suggests that geometric averaging provides no substantial benefit over arithmetic averaging, even when the a_i parameters are quite small—only half as large as the b_i parameters. In real practice data, it is very common for component functions to have an a_i that is large relative to b_i. For example, in the many thousands of practice curves examined in Heathcote et al.'s (2000) survey, 51% yielded parameter estimates for which $a_i > b_i$.

AVERAGING WITHIN A FUNCTION

Among others, K. M. Newell, Liu, and Mayer-Kress (2001) have argued that averaging within curves (e.g., block averaging), like averaging across participants, is dangerous. In their words,

> Learning trials are often blocked . . . to remove the presumed transient randomlike changes from trial to trial while emphasizing the persistent changes or the global trend of learning over trials. The problem is that blocking data from groups of trials can modify or mask properties of the persistent trend as well as those of the transient changes. In particular, this data analysis strategy reduces the evidence of rapid change in performance that is often present early in practice. (p. 59)

Block averaging is an example of the general class of data analysis tools, often called *smooths*, that have been much favored in modern statistics. As its name implies, a smooth is biased when it comes to rapid changes. It is not true, however, that smoothing always changes the shape of *persistent trends*, such as learning curves. We

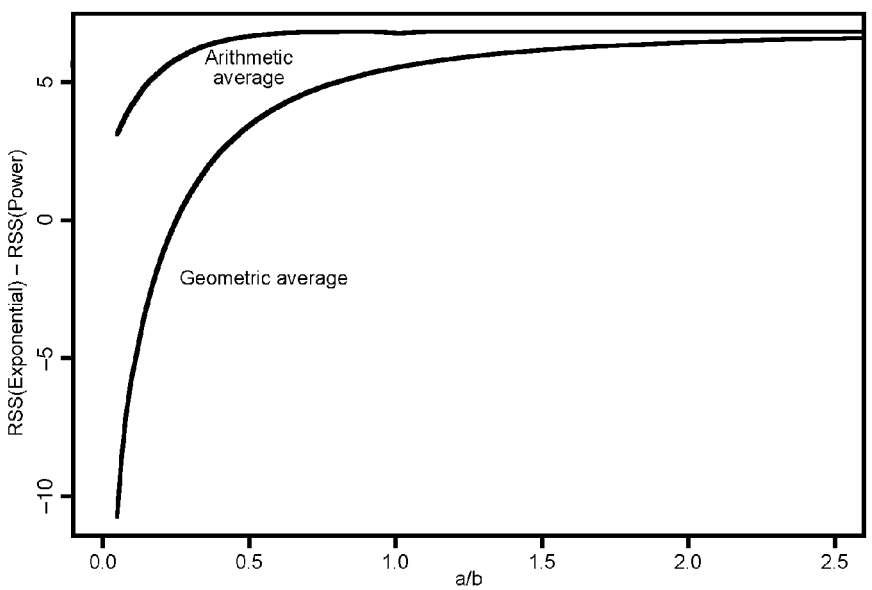

Figure 3. Distortion levels due to arithmetic and geometric averaging of exponential functions. The abscissa represents the ratio of the asymptote parameters of these functions to their scale parameters (a_i/b_i). The ordinate represents the difference between residual sums of squares (RSS) for exponential and power function fits to the average curves. Negative values indicate a better exponential fit, and positive values indicate a better power fit to averaged exponential function data.

will demonstrate that block averaging and more general types of smoothing have *no* effect on the shape of exponential functions and that the bias induced for power functions is usually acceptably small.

Consider the exponential functions defined in Equation 2 and suppose that the covariate x (e.g., practice trials) is measured in N blocks of M trials, so that $x = 1, 2, \ldots, NM$. Each point in the block-averaged series is defined as the arithmetic average of all points within the corresponding block of the raw data series, which yields Equation 3, where i is block number ($i = 1, \ldots, N$):

$$y(i) = a + b \cdot \frac{1}{M} \sum_{j=1}^{M} e^{-r[(i-1)M+j]}. \quad (3)$$

Equation 3 may be reexpressed as Equation 4:

$$y(i) = a + b \cdot \left(\frac{e^{rM}}{M} \sum_{j=1}^{M} e^{-ri} \right) \cdot e^{-rMi}. \quad (4)$$

Thus, the block average is precisely an exponential function, except that the scale parameter (b in Equation 2) is multiplied by a constant and the rate parameter (r in Equation 2) is multiplied by M. The change in scale is linear, and the change in the rate simply reflects a linear change in the units for the predictor (trials to blocks). Hence, there is no distortion of shape.

Similar results hold for moving window smooths, at least when end effects are neglected (end effects can be very complex; see Fan & Gijbels, 1996, or Wand & Jones, 1995). Thus, a simple boxcar smooth—the continuous generalization of block averaging—or a more sophisticated weighted smooth will not change the functional form of exponential functions. Moving window smooths include as a subcase zero-order local polynomial regression (a Nadaraya–Watson smooth), at least for kernels with bounded support. Given the exponential function in Equation 2, the moving window smooth with a kernel of width M is defined as

$$y(x) = a + b \cdot \frac{1}{M} \sum_{j=\frac{M}{2}+1}^{\frac{M}{2}} w_j e^{-r(x+j)}. \quad (5)$$

The w_j are a set of weights (constrained to have sum M), and x is constrained to $(M/2), \ldots, (N-M/2)$, to remove end effects altogether. Equation 5 can similarly be reexpressed as an exponential function in the same form as Equation 2:

$$y(x) = a + b \cdot \left(\frac{1}{M} \sum_{j=\frac{M}{2}+1}^{\frac{M}{2}} w_j e^{-rj} \right) \cdot e^{-rx}. \quad (6)$$

Thus, the weighted moving window smooth of an exponential function is itself an exponential function with the scale parameter multiplied by the term in brackets in Equation 6, while all other parameters remain unchanged.

These results hold only for the data structure described above, with regularly spaced covariate values ($x = 1, 2, \ldots$). This assumption is plausible for learning and memory curves, in which covariate values are most often set by design, but may be unreasonable in other paradigms. If covariate values are subject to independent random variation, the effect on curve form will be small. However, covariate values sometimes vary systematically, with regions of high and low density. In that case, if block averages or smooths are calculated as above (across blocks containing equal numbers of data points), serious distortion can occur: The factor in parentheses in Equation 6 would no longer be constant across the covariate. With systematically variable covariate values, block averaging will be mostly harmless if averages or smooths are instead calculated across fixed widths of the abscissa, so that the number of points in each block varies with covariate density. With that method, the parenthetical term in Equation 6 represents the mean of a variable number of points. It will be relatively constant because those points simply represent more or less dense sampling from the same function (e.g., the mean of {1, 2, 3, 4, 5, 6} is not too different from the mean of {2, 4, 6}).

The power function does not behave quite so tractably under block averaging. In general, it is not the case that the block average of a power function will be a power function itself. However, by applying results from the kernel smoothing literature, we can assure ourselves that the biases introduced by block averaging power function data will be small, given certain conditions. Again, we consider the continuous version of the block average, the boxcar smooth, for generality. Ruppert and Wand (1994) and Bowman and Azzalini (1997) provided estimates for the expected bias of any kernel smooth, including the boxcar. Assuming that the "true" regression function is a power function, the approximate (first order) pointwise bias for the boxcar smooth is $\frac{1}{48} M^2 bs(s+1)x^{-s-2}$. As K. M. Newell et al. (2001) anticipated, the greatest bias occurs at the start of the series, because that is where curvature is greatest, but it rapidly diminishes to zero toward the tail of the series (as $x \to \infty$). This bias will be small relative to the criterion values, as long as small enough values of M are chosen. For the particular case of power-exponential comparison, the block average of a power function will be, if anything, less like an exponential function than the raw data.[4]

Reanalyses of Published Data

Figure 4 shows the effect of block averaging on exponential versus power function discrimination in practice data analyzed by Heathcote et al. (2000). Each of the 17 unaveraged data sets from the practice law survey was reanalyzed after averaging over blocks. Both short and long block averages were used for each individual series, with the actual block lengths dictated by the design of the experiment and ensuring that each series had a reasonable number of points after averaging (see Appendix C for details).

The effect of block averaging was much smaller than the previously described effect of averaging over participants. In particular, preference for the exponential

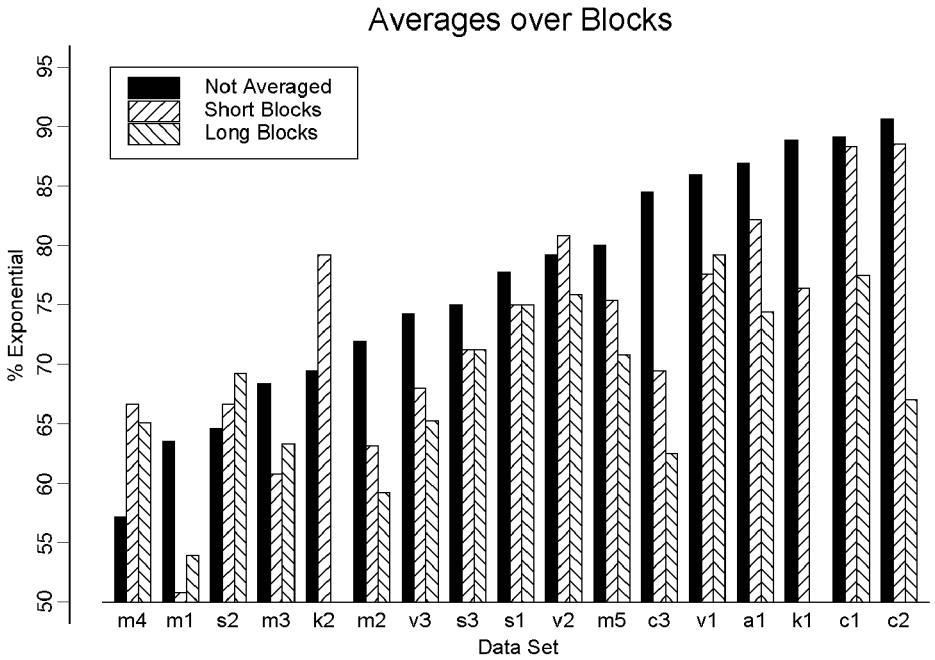

Figure 4. Percentages, over conditions and participants, of fits where a three-parameter exponential function provided a better fit than did a three-parameter power function in unaveraged data, and corresponding percentages, over conditions only, for fits to short and long block arithmetic averages. Note that only short blocks were used for the *k*1 and *k*2 data sets, owing to the design of the experiment.

model was never reversed, although preference for the exponential was generally less than that for the unaveraged case. Usually, the increased averaging associated with longer blocks resulted in a greater decrease in preference for the exponential, as compared with shorter blocks. The exceptions to these generalizations occurred mainly in data sets with weaker unaveraged preference, where block averaging sometimes increased preference for the exponential and in one case (s2) long blocks caused a greater increase than short blocks. Averages over participants were also calculated on block averages. The effect on model selection was similar to that seen with the previous averages over participants (Figure 2)—that is, large distortions were observed.

DISCUSSION

Arithmetic averaging preserves the functional form of its components only under very strict conditions and can have opposing effects on averages of noisy exponential functions. Where the individual curves' rate parameters vary sufficiently, a bias toward the power function is created in the average. Averaging can also reduce noise, so in some cases, a clearer preference for the exponential can emerge as the deleterious effects of noise on model discrimination are attenuated. In the data from Heathcote et al.'s (2000) survey (Figure 2, above), these effects appear to interact in complicated ways. Consequently, researchers who rely on the average could be misled into concluding that different functional forms apply to different conditions or paradigms, when the real cause is variation in the distribution of exponential learning rates.

Geometric averaging cannot be relied upon to cure averaging distortion for power and exponential functions. Geometric averaging attenuates the bias favoring the power function over the exponential function only when asymptotic performance is less than one third of the scale of learning. For practice curves, this condition is commonly violated. Reanalysis of the practice data analyzed by Heathcote et al. (2000) showed that geometric averaging did not differ much from arithmetic averaging and, so, was largely ineffective at avoiding distortion.

A much more optimistic picture emerged for block averages and, in general, for *smooths* that aggregate data from contiguous trials. K. M. Newell et al.'s (2001) concern about block averaging is unwarranted for exponential learning curves, and the bias for power functions is generally small for reasonable choices of smoothing parameters (e.g., block width). A comparison of Figures 2 and 4 shows that, for practice data, block averages produced much less distortion than did averages over participants.

All of the types of averaging examined here improved the goodness of fit for both exponential and power functions for the data of Heathcote et al.'s (2000) survey. For the raw data, the (unweighted) average R^2 across data sets

was .35 for the exponential function and .31 for the power function. For short block averages, the average R^2 was .61 for the exponential function and .57 for the power function, and for long block averages the average R^2 was .73 for the exponential function and .70 for the power function. In arithmetic averages over participants, the average R^2 was .70 for the exponential function and .65 for the power function. Hence, block averaging can be as effective in improving signal-to-noise ratio as averaging over participants. Consequently, block averaging can take advantage of the improvement in model discrimination that occurs with decreased noise, while introducing no averaging distortion for the exponential function and very little averaging distortion for the power function.

When we have reported our results on the potential distortion due to averaging over participants to colleagues, one of the first questions to arise regards the implications for the analysis of learning curves with a repeated measures analysis of variance (ANOVA). The answer to this question is complicated, and we can deal with it only briefly here. However, it can be definitely stated that the object of inference in an ANOVA is the mean over participants, so any quantitative evaluation of the mean function's shape, such as polynomial contrasts, can suffer from averaging distortion.

The situation is often much worse than it need be: Simple additivity, rather than linearity, is usually adopted as the structural model for the subjects' effect in most repeated measures ANOVA programs. Additivity implies that each participant differs in location (e.g., asymptote) but exhibits the same change in performance from the beginning to the end of learning, an erroneous assumption in our experience. As is shown in Appendix A, no averaging distortion occurs when participants' curves differ in scale as well as in location; it seems wasteful not to take advantage of this fact. When only additivity is assumed, scale differences between participants are assigned to error, unnecessarily reducing the power of tests. Scale differences can also induce spurious covariance between levels of the learning factor, which are commonly corrected by reducing degrees of freedom, at a further unnecessary cost to power. Mandel (1963) provides the methods necessary to perform an ANOVA allowing a full linear model for subjects' effects. Heathcote, Mewhort, and Brown (2002) have extended this approach to allow violations of the linear subjects' effect model and, hence, the potential for averaging distortion, to be detected.

The results presented here have implications for theories of learning, as well as for analysis of learning curves. In many cases, measurement and design limitations mean that individual participants' data are the result of a summation or averaging process. In retention experiments, for example, individual participant retention probabilities are commonly obtained by averaging over a population of items. If items have widely differing exponential rates of forgetting, the *individual* retention curve can appear to have a power form (see Heathcote et al., 2000, for a discussion of this issue for practice theories). The present results show that exponential component rates need differ only by an order of magnitude for a power function to provide a better fit to the average. In general, if a theory postulates that observed performance is the result of the summation or averaging of components that differ nonlinearly, the effects of summation or averaging must be taken into account in determining the theories' predictions for performance. A. Newell and Rosenbloom (1981) acknowledged just this possibility, and Neves and Anderson (1981) provided such a theory, in which a supposed power function for individual practice curves is explained as resulting from the sum of a series of stages that learn exponentially.

Our analyses focused on exponential functions because they are arguably the simplest plausible form for a learning curve; their shape is defined by a single nonlinear rate parameter, and any unobserved learning trials prior to an experiment (k) can always be absorbed into a linear scale parameter (e.g., $e^{r(x+k)} = e^{rk}e^{rx}$).[5] Given that averages of exponential functions demonstrably produce marked distortions, the situation will likely be worse for more complex nonlinear models. For power learning, for example, prior practice introduces a second nonlinear parameter, so variations in prior practice among participants can increase averaging distortion. In general, therefore, we concur with Massaro's (1998) statement that if only averages are examined, "we might have an explanation for an average subject, but one that does not apply to any of the actual individuals making up the average. Thus averaging may preclude the discovery of important properties" (p. 132).

REFERENCES

Aczel, J. (1966). *Lectures on functional equations and their applications.* London: Academic Press.

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

Bahrick, H. P., Fitts, P. M., & Briggs, G. E. (1957). Learning curves—facts or artifacts? *Psychological Bulletin*, **54**, 256-268.

Bakan, D. (1954). A generalization of Sidman's results on group and individual functions and a criterion. *Psychological Bulletin*, **51**, 63-64.

Becker, R. A., Chambers, J. M., & Wilks, A. R. (1988). *The new S language.* Pacific Grove, CA: Wadsworth & Brooks.

Boas, F. (1892). The growth of children. *Science*, **19**, 256-257, 281-282; **20**, 351-352.

Bowman, A. W., & Azzalini, A. (1997). *Applied smoothing techniques for data analysis: The kernel approach with S-Plus illustrations.* Oxford: Oxford University Press, Clarendon Press.

Brown, S., & Heathcote, A. (in press). Bias in exponential and power function fits due to noise: Comment on Myung, Kim, and Pitt. *Memory & Cognition.*

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.

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

ESTES, W. K. (2002). Traps in the route to models of memory and decision. *Psychonomic Bulletin & Review*, **9**, 3-25.
FAN, J., & GIJBELS, I. (1996). *Local polynomial modelling and its applications*. London: Chapman & Hall.
HEATHCOTE, A., BROWN, S., & MEWHORT, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, **7**, 185-207.
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., & BROWN, S. (2002). *Average curves and repeated-measures analysis of variance*. Manuscript in preparation.
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.
MANDEL, J. (1963). Non-additivity in two-way analysis of variance. *Journal of the American Statistical Society*, **56**, 878-888.
MASSARO, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.
MYUNG, I. J., KIM, C., & PITT, M. A. (2000). Toward an explanation of the power law artifact: Insights from response surface analysis. *Memory & Cognition*, **28**, 832-840.
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., & 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.
NEWELL, K. M., LIU, Y.-T., & MAYER-KRESS, G. (2001). Time scales in motor learning and development. *Psychological Review*, **108**, 57-82.
PALMERI, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 324-354.
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.
RUPPERT, D., & WAND, M. P. (1994). Multivariate locally weighted least squares regression. *Annals of Statistics*, **22**, 1347-1370.
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.
SIDMAN, M. (1952). A note on functional relations obtained from group data. *Psychological Bulletin*, **49**, 263-269.
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.
THOMAS, E. A. C., & ROSS, B. H. (1980). On appropriate procedures for combining probability distributions within the same family. *Journal of Mathematical Psychology*, **21**, 136-152.
UNDERWOOD, B. (1949). *Experimental psychology*. New York: Appleton-Century-Crofts.
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.
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. The geometric average of $\{x_i: i = 1 \ldots n\}$ is

$$\exp\left(\frac{1}{n}\sum_{i=1}^{n}\ln(x_i)\right).$$

2. Logarithmic (rather than linear) spacing was used in keeping with the ratio definition of *r*-spacing used above.

3. The asymptote estimates *mean* performance after extended practice. Consequently it is larger than many observed values later in practice.

4. For $y = bx^{-s}$ and bias $= kbx^{-(s+2)} + O(x^{-(s+3)})$, where $k = M^2 s(s+1)/48$, the approximate block average power function is $z = y +$ bias. The derivative of z, neglecting higher order terms, is

$$z' = -sx^{-1}z - 2kbx^{-(s+3)},$$

and so the approximate negative logarithmic derivative of z is

$$-\frac{z'}{z} = \frac{s}{x} + \frac{2k}{x\left(x^2+k\right)},$$

which decreases faster than a hyperbolic function (the relative rate of a power function) as $s > 0$ and, so, $k > 0$. For an exponential function, in contrast, the relative rate is constant. See Heathcote et al. (2000) for more details on relative learning rates.

5. Note that it is this translation invariance of shape that allows local averages to exactly preserve the form of exponential functions.

(*Continued on next page*)

APPENDIX A
Representative Arithmetic Averages

THEOREM. *An arithmetic average of several component functions will have the same form as those components if and only if the component functions are linear in all parameters that vary across components.*

PROOF. A simple proof is presented below, but its essential idea has been known since at least 1821, when Cauchy published it; the interested reader is referred to Aczel (1966) for extensions (e.g., averages other than arithmetic). Consider a set of P component dependent variables, y, that are all a real-valued function dependent on different k-dimensional parameter vectors, A_i and a vector of covariates, x (fixed for all i). To prove sufficiency, assume that the functions, y, are linear in the parameter vectors, A_i—that is,

$$y(A_i + A_j, \boldsymbol{x}) = y(A_j, \boldsymbol{x}) + y(A_j, \boldsymbol{x}) \quad \forall A_i, A_j \in \boldsymbol{R}^k$$
$$y(KA_i, \boldsymbol{x}) = K \cdot y(A_i, \boldsymbol{x}) \qquad \forall A_i \in \boldsymbol{R}^k, \qquad K \in \boldsymbol{R}.$$

Note that the second property follows from the first, at least for rational K. The same is true for all real K by the usual limit argument, if y is continuous. Under these conditions the arithmetic average (AA) of the y over the P parameter vectors will be representative:

$$AA[y(A_i, \boldsymbol{x})] = \frac{1}{P}\sum_{i=1}^{P} y(A_i, \boldsymbol{x}) = \frac{1}{P} y\left(\sum_{i=1}^{P} A_i, \boldsymbol{x}\right)$$
$$= y\left(\frac{1}{P}\sum_{i=1}^{P} A_i, \boldsymbol{x}\right) = y[AA(A_i), \boldsymbol{x}].$$

To prove necessity, note that the above chain of implications runs backward also: The third expression follows from the fourth by invoking the additivity assumption, and the second follows from the third by invoking the scalar multiplication assumption (or the additivity assumption, if y is continuous).

APPENDIX B
Computational Details

All simulations used scripts written by the authors in the S language (Becker, Chambers, & Wilks, 1988). Validation of nonlinear regression solutions was made using a program written by the authors in Pascal. The primary nonlinear regression algorithm performed least-squares minimization by implementing a quasi-newton algorithm and an iterative approximation to the Hessian matrix. Start points for searches for exponential and power function parameters were generated automatically by heuristics based on approximating the asymptotes and then estimating the other parameters by linear regression, followed by grid search around the estimates for robustness. Minimizations to ascertain the distortion ratio were carried out using a golden-section line search, with starting points generated from the outputs of neighboring searches. When $a_i = 0$ for all i, linear least-squares regression after log-log transformation was used to determine the best-fitting power function, and linear least-squares regression after log-linear transformation was used to determine the best-fitting exponential function.

Four different values of M (30, 100, 300, and 1,000) were used. At each value of M, 10 values of r_{max} were used. The values of r_{max} were chosen from the feasible range—those for which distortion was possible. Choosing values of r_{max} for which no distortion was possible resulted in false convergence estimates in the search for the distortion ratio (since the objective function was relatively constant across different ratios). The r_{max} values were chosen so that r_{max} M ranged from approximately 100 down to 3 in logarithmically spaced steps.

In Simulations 2 and 3, different numbers of curves were also averaged—2, 3, 4, 5, 7, 9, 12, 15, or 20—representing averaging across different numbers of participants. A constant value of r_{max} was used for each curve in the average, and values of r below r_{max} were logarithmically spaced between r_{max} and r_{min}.

APPENDIX C
Reanalyzed Data Sets

Table C1 defines the labels used for data sets from Heathcote et al.'s (2000) survey. The "Short Blocks" and "Long Blocks" columns indicate the number of observations per block/number of blocks per participant. For the k1 and k2 data, different block lengths were used for the two within-subjects conditions.

Ordinary least-squares estimation was used to fit three-parameter power and exponential functions with estimated asymptotes bounded below by zero. Note that the preaveraging results in Figure 2 differ slightly from those reported in Heathcote et al. (2000), for two reasons. First, the numbering system used for the practice factor (N) labeled the first correctly answered trial occurring in each within-subjects condition as 1, the second correct trial as 2, and so on, rather than using the absolute trial number regardless of condition, as in Heathcote et al. Second, practice series were truncated to the length of the shortest practice series within a condition. This ensured that each participant contributed exactly one RT to each value in the averaged series. We found that other numbering systems that did not enforce this condition introduced substantial distortion into the average—for instance, by allowing the tail of the series to be dominated by a single participant's data. The disadvantage of this approach is that it discards some information about the tail of the practice function and, so, may push the results toward no preference (50%). A comparison with Heathcote et al.'s Figure 1 shows that the effect of these changes was only slight. Note that all data sets that were fit separately for different response strategies in Heathcote et al. were also fit separately here. Grouped results are reported, treating algorithm versus memory strategies as an extra within-subjects condition.

Table C1
Labels, Sources, and Block Size/Number of Blocks for the 17 Unaveraged Practice Data Sets From Heathcote, Brown, and Mewhort (2000)

Label	Reference	Experiment	Long Blocks	Short Blocks
m1	Rickard & Bourne (1996)	"OPER" Experiment	10/9	3/30
m2	Rickard (1997)	"CPL" Experiment	10/9	5/18
m3	Reder & Ritter (1992); also Delaney, Reder, Staszewski, & Ritter (1998)	Experiment 1	4/5	2/10
m4	As for m3	Experiment 2	4/5	2/10
m5	Schunn, Reder, Nhouyvanisvong, Richards, & Stroffolino (1997); also Delaney, Reder, Staszewski, & Ritter (1998)	Experiment 1, using only stimuli presented 28 times.	4/7	2/14
s1	Strayer & Kramer (1994b)	consistently mapped trials from mixed consistent/varied mapping training blocks from Experiment 2	48/15	24/30
s2	Strayer & Kramer (1994b, 1994c)	consistently mapped training blocks from Experiment 2 of 1994a and Experiments 4, 6, and 7 from 1994b and from an unpublished two-alternative forced-choice version of the task	48/15	24/30
s3	Strayer & Kramer (1994a)	consistently mapped trials (young participants)	24/18	12/36
v1	Heathcote & Mewhort (1993)	Experiment 1	20/10	10/20
v2	Carrasco, Ponte, Rechea, & Sampedro (1998)		12/7	4/21
v3	As for v1	Experiments 3 and 4	20/16	10/32
k1	Verwey (1996)	time to press each key taken separately—Day 1 session only	–	30/24 & 10/12
k2	As for k1	time to press each key taken separately—Day 1 session omitted due to nonstationary errors	–	30/53 & 10/21
c1	Palmeri (1997)	Experiment 1	16/13	4/52
c2	As for c1	Experiment 2	16/10	4/40
c3	As for c1	Experiment 3	8/10	4/20
a1	As for m2	alphabet arithmetic task	12/7	4/21

(Manuscript received February 26, 2002;
revision accepted for publication August 13, 2002.)