

# **Averaging Learning Curves Across and Within Participants**

Scott Brown and Andrew Heathcote

School of Behavioural Sciences,  
The University of Newcastle, Australia

Address for correspondence  
Scott Brown

School of Behavioural Sciences, Building W  
The University of Newcastle, Callaghan, 2308, NSW, Australia

Email: [scott@baal.newcastle.edu.au](mailto:scott@baal.newcastle.edu.au)

## **Abstract**

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 averaging distortion.

Averaging is perhaps the most common statistical analysis technique used in psychological research today. Researchers average data 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 behaviour 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 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 (e.g., Bakan, 1954; Bahrick, Fitts & Briggs, 1957; 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 couldn't be exactly exponential – does not rule out the possibility the average might be close enough to exponential for practical purposes.

Two surveys of skill acquisition data (Heathcote, Brown & Mewhort, 2000; Newell & Rosenbloom, 1981) have provided evidence that distortions due to averaging have led to misguided theory development in psychology. In this paper we endeavour to augment the averaging literature by providing several results dealing with averages across different curves: (a) a simplified version of an established proof outlining exactly when arithmetic averages of any nonlinear function will be distortion-free; (b) Monte-Carlo simulations showing that averages of exponential functions can easily mislead model discriminations; and (c) re-analyses 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: (c) that there are methods of averaging within curves that can result in little or no distortion, but still provide strong noise reduction benefits; and (e) 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 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 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) define 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 only makes sense 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 only makes sense 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,  $\mathbf{y}: \mathbb{R}^k \rightarrow \mathbb{R}$ , this means that if  $\mathbf{y}$  is to have a representative average, then it can be expressed as the inner product of a fixed parameter vector  $\mathbf{A}$  with a  $\mathbf{k}$ -vector-valued function of  $\mathbf{x}$ ,  $\mathbf{f}(\mathbf{x}, \boldsymbol{\theta})$  where  $\boldsymbol{\theta}$  is a fixed vector of non-linear parameters. So, any

function for which the arithmetic average will be representative may be expressed in the form:

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

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 ( $\theta$ ) 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 Tweeny (1997), Wixted and Ebbesen (1997) report little averaging distortion in the data of Wixted and Ebbesen (1991), as both individual and average retention functions from were better fit by power than exponential functions. For averaged practice curves, 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, analysed 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 aimed at reconciling 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. 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,  $\underline{s}$  for the power and  $\underline{r}$  for the exponential:

$$y_P = a_p + b_p x^{-s} \quad (1)$$

$$y_E = a_E + b_E e^{-rx} \quad (2)$$

Anderson and Tweney (1997) examined only two-parameter versions of the power and exponential functions, with  $\underline{a}$  constrained to be zero in both cases. In the two-parameter case, geometric averaging<sup>1</sup> provides an easy solution to averaging distortion – but one that cannot be extended to other functions in which the asymptote ( $\underline{a}$ ) parameters vary. Our simulations extend Anderson and Tweney’s work to the three-parameter case, use 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  $\underline{x}=1, 2, \dots, \underline{M}$  indexes the values of the covariate (e.g., experimental trials), and  $\underline{i}=1, 2, \dots, \underline{P}$  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  $\underline{a}_i=0$ , and determined whether the average curve was better described by an exponential or power function. The spread of the component functions’  $\underline{r}$  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  $\underline{r}$  values,  $\underline{r}_{\max}$ , was fixed and a line search over the other  $\underline{r}$  parameter,  $\underline{r}_{\min}$ , 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  $\underline{r}_{\max}:\underline{r}_{\min}$  at the point where the average curve changed from more like an exponential 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  $\underline{r}_{\max}$  and of  $\underline{M}$ , although only two component functions were used in Simulation 1 ( $\underline{P}=2$ ). The  $\underline{b}_i$  parameters were constrained to be constant across component functions.

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

Simulation 3 extended Simulation 2 by using component functions with values of  $\underline{a}_i > 0$  and values of  $\underline{b}_i$  that varied across component functions (i.e., three parameter functions). These simulations were designed to check whether using constant (and zero) asymptotes and constant  $\underline{b}_i$  values had restricted the results of the previous simulations. Analysis was complicated by the non-zero  $\underline{a}_i$  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  $\underline{a}_i > 0$  and varying  $\underline{b}_i$  yielded the same outcome as the simulations with  $\underline{a}_i = 0$  and  $\underline{b}_i$  fixed. This result concurs with Anderson and Tweney (1997), who found that varying  $\underline{b}_i$  values had very little effect on averaging artefacts (they did not vary  $\underline{a}_i$ ). 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 lognormal errors for the transformed linear regressions.

## Simulation Results

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

-----  
 Insert Figure 1 about here  
 -----

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 behaviour 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 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.

### ***Re-Analyses of Published Data Sets***

To examine the impact of averaging on real psychological data, 17 data sets analysed by Heathcote et al. (2000) were re-analysed 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 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 data sets showed a consistent preference for the exponential function (in 62% to 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; 2 cases resulted in a strong preference for the power function.

-----  
Insert Figure 2 about here  
-----

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 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 (accepted) 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 analysing 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, due to noise<sup>3</sup>, in which case a logarithm is not defined.

However, even when non-zero 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  $\underline{a}_i = 0$ , geometric averaging completely removed the bias towards the power function, as expected. As  $\underline{a}_i$  increased this benefit decreased. These effects are illustrated in Figure 3 as a function of the ratio  $\underline{a}_i/\underline{b}_i$ . 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.

-----  
 Insert Figure 3 about here  
 -----

Arithmetic averaging, for all values of  $\underline{a}_i/\underline{b}_i$ , causes distortion towards the power function. For values of  $\underline{a}_i/\underline{b}_i$  less than about 0.3, geometric averaging was relatively effective in removing bias towards the power function. At all values of  $\underline{a}_i/\underline{b}_i$  greater than this, the bias towards the power function was strong enough to reverse model discrimination choices made on the basis of RSS. For values of  $\underline{a}_i/\underline{b}_i$  greater than approximately 0.5, the bias due to geometric averaging is 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  $\underline{a}_i$  parameters are quite small: only half as large as the  $\underline{b}_i$  parameters. In real practice data, it is very common for component functions to have  $\underline{a}_i$  large relative to  $\underline{b}_i$ . For example, in the many thousands of practice curves examined in Heathcote et al.'s (2000) survey, 51% yielded parameter estimates for which  $\underline{a}_i > \underline{b}_i$ .

## **Averaging Within a Function**

Among others, 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 favoured 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 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  $\underline{x}$  (e.g., practice trials) is measured in  $\underline{N}$  blocks of  $\underline{M}$  trials, so that  $\underline{x} = 1, 2, \dots, \underline{NM}$ . 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  $\underline{i}$  is block number ( $\underline{i} = 1, \dots, \underline{N}$ ):

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

Equation 3 may be re-expressed as Equation 4:

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

Thus, the block average is precisely an exponential function, except the scale parameter ( $\underline{b}$  in Equation 2) is multiplied by a constant, and the rate parameter ( $\underline{r}$  in Equation 2) is multiplied by  $\underline{M}$ . 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 Wand & Jones, 1995, or Fan & Gijbels, 1996). Thus, a simple boxcar smooth – the continuous generalisation of block averaging – or a more sophisticated weighted smooth will not change the

functional form of exponential functions. Moving window smooths include as a sub-case 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  $\underline{M}$  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  $\underline{M}$ ) and  $\underline{x}$  is constrained to  $(\underline{M}/2), \dots, (\underline{N}-\underline{M}/2)$ , to remove end effects altogether. Equation 5 can similarly be re-expressed 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 ( $\underline{x} = 1, 2, \dots$ ). 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) provide 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  $\frac{1}{48} M^2 b_s (s+1) x^{-s-2}$ . As 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 towards the tail of the series (as  $\underline{x} \rightarrow \infty$ ). This bias will be small relative to the criterion values, as long as small enough values of  $\underline{M}$  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<sup>4</sup>.

### ***Re-Analyses of Published Data***

Figure 4 shows the effect of block averaging on exponential versus power function discrimination in practice data analysed by Heathcote et al. (2000). Each of

the 17 un-averaged data sets from the practice law survey was re-analysed 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).

-----  
Insert Figure 4 about here  
-----

The effect of block averaging was much smaller than the previously described effect of averaging over participants. In particular, preference for the exponential model was never reversed, although preference for the exponential was generally less than for the un-averaged case. Usually, the increased averaging associated with longer blocks resulted in a greater decrease in preference for the exponential compared to shorter blocks. The exceptions to these generalizations occurred mainly in data sets with weaker un-averaged 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 favouring the power over exponential functions only when asymptotic performance is less than one third of the scale of learning. For practice curves, this condition is commonly violated. Re-analysis of the practice data analysed by Heathcote et al. (2000) showed geometric averaging did not differ much from arithmetic averaging, and so it 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. 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). Comparison of Figures 2 and 4 show that, for practice data, block averages produced much less distortion than 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  $\underline{R}^2$  across data sets was .35 for the exponential and .31 for the power function. For short block averages, the average  $\underline{R}^2$  was .61 for the exponential and .57 for the power function, and for long block averages the average  $\underline{R}^2$  was .73 for the exponential function and .70 for the power

function. In arithmetic averages over participants the average  $\underline{R}^2$  was .70 for the exponential 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, 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 repeated measures 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 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 end of learning, an erroneous assumption in our experience. As shown in Appendix A, no averaging distortion occurs when participants' curves differ in scale as well as 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 ANOVA allowing a full linear model for subjects' effects. Heathcote, Mewhort and Brown (in preparation) 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 participant's 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. Newell and Rosenbloom (1981) acknowledge just this possibility, and Neves and Anderson (1981) provide 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.,  $e^{r(x+k)} = e^{rk} e^{rx}$ )<sup>5</sup>. 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 (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.
- Bahrack, 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, California: 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: Clarendon Press.
- Brown, S. & Heathcote, A. (accepted) 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. 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, K. W. (1956). The problem of inference from curves based on group data. Psychological Bulletin, 53, 134-140.

Estes, K. W. (2002). Traps in the route to models of memory and decision. Psychonomic Bulletin & Review, *9*(1) 3-26

Fan, J. & Gijbels, I. (1996) Local polynomial modelling and its applications. London: Chapman & Hall.

Heathcote, A., Brown, S. & Mewhort, D.J.K. (2000) Repealing the power law: The case for an exponential law of practice. Psychonomic Bulletin & Review, *7*, 185-207.

Heathcote, A., Mewhort, D.J.K. & Brown, S. (in preparation). Average curves and repeated-measures analysis of variance.

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

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, Mass.: MIT Press.

Myung, I. J., Kim, C., & Pitt, M. A. (2000). Toward an explanation of the power-law artefact: Insights from response surface analysis. Memory & Cognition, *28*, 832-840.

Newell, K. M., Liu, Y-T & Mayer-Kress, G. (2001) Time scales in motor learning and development. Psychological Review, *108*, 57-82.

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.

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.

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). Strategies and automaticity I. Journal of Experimental Psychology: Learning Memory & Cognition, 20, 318-341.

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

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

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(5) 731-739.

## 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 it's 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  $\underline{P}$  component dependent variables,  $\underline{y}$ , that are all a real-valued function dependent on different  $\underline{k}$ -dimensional parameter vectors,  $\underline{A}_i$  and a vector of covariates,  $\underline{x}$  (fixed for all  $i$ ). To prove sufficiency, assume that the functions,  $\underline{y}$ , are linear in the parameter vectors,  $\underline{A}_i$ , that is:

$$\begin{aligned} y(A_i + A_j, \mathbf{x}) &= y(A_i, \mathbf{x}) + y(A_j, \mathbf{x}) \quad \forall A_i, A_j \in \mathbf{R}^k \\ y(KA_i, \mathbf{x}) &= K \cdot y(A_i, \mathbf{x}) \quad \forall A_i \in \mathbf{R}^k, \quad K \in \mathbf{R} \end{aligned}$$

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

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

To prove necessity, notice that the above chain of implications runs backwards 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 the  $\underline{y}$  are 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 minimisation 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 asymptotes and then estimating the other parameters by linear regression, followed by grid-search around the estimates for robustness. Minimisations to ascertain the distortion ratio were carried out using a golden-section line search, with starting points generated from the outputs of neighbouring searches. When  $\underline{a}_i=0$  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  $\underline{M}$  (30, 100, 300 and 1000) were used. At each value of  $\underline{M}$ , 10 values of  $\underline{r}_{\max}$  were used. The values of  $\underline{r}_{\max}$  were chosen from the feasible range – those for which distortion was possible. Choosing values of  $\underline{r}_{\max}$  for which no distortion was possible resulted in false convergence estimates in the search for the distortion ratio (as the objective function was relatively constant across different ratios). The  $\underline{r}_{\max}$  values were chosen such that  $\underline{r}_{\max} \underline{M}$  ranged from approximately 100 down to three in logarithmically spaced steps.

In simulations 2 and 3, different numbers of curves were also averaged – either 2, 3, 4, 5, 7, 9, 12, 15 or 20 – representing averaging across different numbers

of participants. A constant value of  $\underline{r}_{\max}$  was used for each curve in the average and values of  $\underline{r}$  below  $\underline{r}_{\max}$  were logarithmically spaced between  $\underline{r}_{\max}$  and  $\underline{r}_{\min}$ .

## Appendix C: Re-Analysed Data Sets

Table A1 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-subject conditions.

-----

Insert Table A1 about here

-----

Ordinary least squares estimation was used to fit three parameter power and exponential functions with estimated asymptotes bounded below by zero. Note that the pre-averaging 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) labelled 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 towards no preference (50%). 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 vs. memory strategies as an extra within-subjects condition.

## Tables

Table A1: Labels, sources and block size/number of blocks for the 17 un-averaged practice data sets from Heathcote et al. (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 (1994a)	Consistently mapped trials from mixed consistent/varied mapping training blocks from Experiment 2	48/15	24/30
s2	Strayer & Kramer (1994a; 1994b)	Consistently mapped training blocks from Experiment 2 of (1994a), and Experiments 4, 6, and 7 from (1994b), and an unpublished two-alternative forced-choice version of the task	48/15	24/30
s3	Strayer & Kramer (1994c)	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

<b>k2</b>	<b>As for k1</b>	<b>Time to press each key taken separately – day 1 session omitted due to non-stationary errors.</b>	-	30/53 & 10/21
<b>c1</b>	<b>Palmeri (1997).</b>	<b>Experiment 1</b>	16/13	4/52
<b>c2</b>	<b>As for c1</b>	<b>Experiment 2</b>	16/10	4/40
<b>c3</b>	<b>As for c1</b>	<b>Experiment 3</b>	8/10	4/20
<b>a1</b>	<b>As for m2</b>	<b>Alphabet Arithmetic task</b>	12/7	4/21

---

## Figure Captions

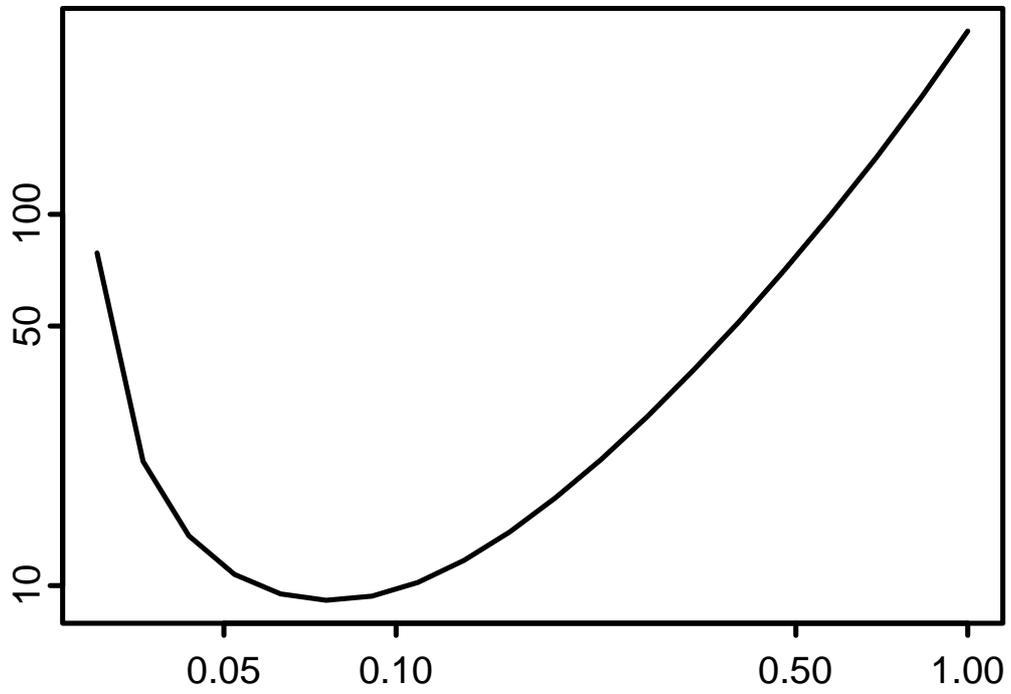
Figure 1. Log-log plot of distortion ratio vs.  $r_{\max}$  for a typical function set (two functions, length 100,  $\underline{a}_1 = \underline{a}_2 = 0$ ,  $\underline{b}_1 = \underline{b}_2 = 5000$ ).

Figure 2. Percentages, over conditions and participants, of fits where a three-parameter exponential function provided a better fit than a three-parameter power function in un-averaged 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.

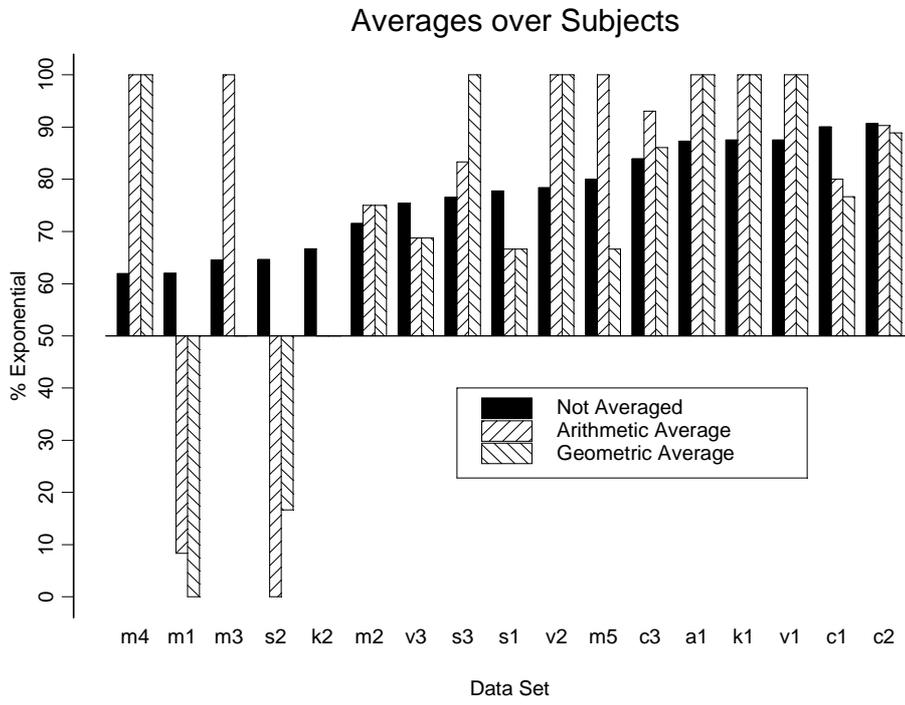
Figure 3. Distortion levels due to arithmetic and geometric averaging of exponential functions. The abscissa represents the ratio of the asymptote parameters of these functions to their scale parameters ( $\underline{a}_i/\underline{b}_i$ ). The ordinate represents the difference between residual sums 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.

Figure 4. Percentages, over conditions and participants, of fits where a three-parameter exponential function provided a better fit than a three-parameter power function in un-averaged 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 k1 and k2 data sets due to the design of the experiment.

**Figure 1**



**Figure 2**



**Figure 3**

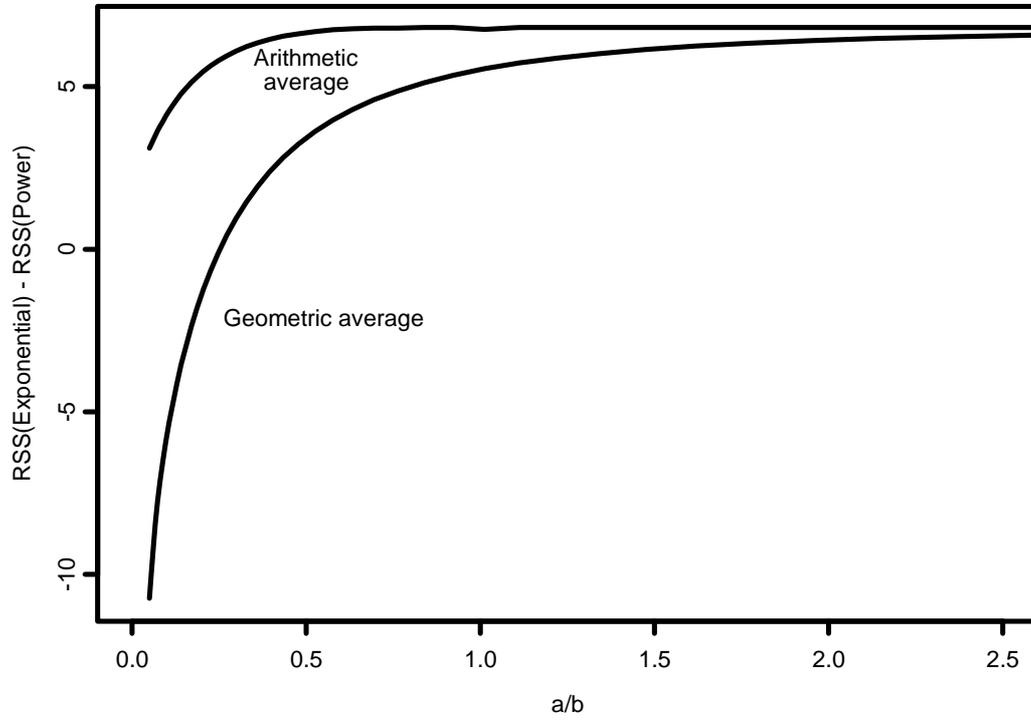
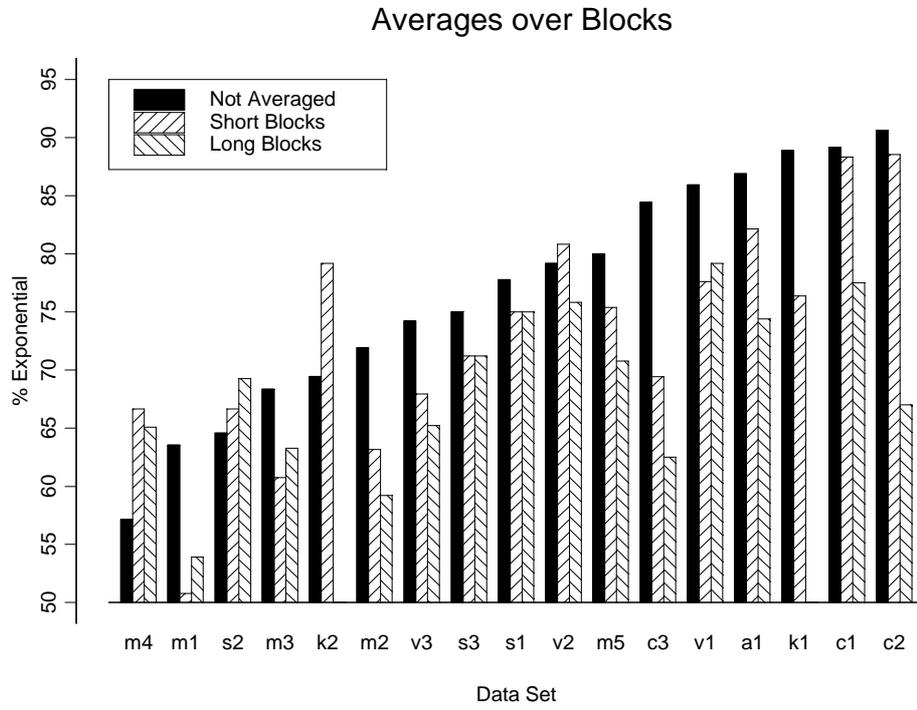


Figure 4



## Footnotes

---

<sup>1</sup> The geometric average of  $\{\underline{x}_i; i=1..n\}$  is  $\exp\left(\frac{1}{n} \sum_{i=1}^n \ln(x_i)\right)$ .

<sup>2</sup> Logarithmic (rather than linear) spacing was used in keeping with the ratio definition of  $\underline{r}$ -spacing used above.

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

<sup>4</sup> For  $\underline{y} = \underline{b}\underline{x}^{-\underline{s}}$  and bias =  $\underline{k}\underline{b}\underline{x}^{-(\underline{s}+2)} + O(\underline{x}^{-(\underline{s}+3)})$ , where  $\underline{k} = \underline{M}^2 \underline{s}(\underline{s}+1)/48$ , the approximate block average power function is  $\underline{z} = \underline{y} + \text{bias}$ . The derivative of  $\underline{z}$ , neglecting higher order terms, is  $z' = -s\underline{x}^{-1}z - 2k\underline{b}\underline{x}^{-(\underline{s}+3)}$ , and so the approximate negative logarithmic derivative of  $\underline{z}$  is  $-\frac{z'}{z} = \frac{s}{x} + \frac{2k}{x(x^2 + k)}$ , which decreases faster than a hyperbolic

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

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