*Memory & Cognition*
*2003, 31 (4), 656-661*

# Notes and Comment

## Bias in exponential and power function fits due to noise: Comment on Myung, Kim, and Pitt

SCOTT BROWN
*University of California, Irvine, California*

and

ANDREW HEATHCOTE
*University of Newcastle, Callaghan, New South Wales, Australia*

*Myung, Kim, and Pitt (2000) demonstrated that simple power functions almost always provide a better fit to purely random data than do simple exponential functions. This result has important implications, because it suggests that high noise levels, which are common in psychological experiments, may cause a bias favoring power functions. We replicate their result and extend it by showing strong bias for more realistic sample sizes. We also show that biases occur for data that contain both random and systematic components, as may be expected in real data. We then demonstrate that these biases disappear for two- or three-parameter functions that include linear parameters (in at least one parameterization). Our results suggest that one should exercise caution when proposing simple power and exponential functions as models of learning. More generally, our results suggest that linear parameters should be estimated rather than fixed when one is comparing the fit of nonlinear models to noisy data.*

Learning performance ($y$) is usually modeled as the sum of a smooth deterministic function ($f$) of a predictor ($t$, e.g., time or practice trials) and a noise term ($\varepsilon$). Recently, Myung, Kim, and Pitt (2000) reported that a simple power function ($y = t^{\beta}$) almost always provides a better fit to purely random data than does a simple exponential function ($y = e^{\alpha t}$). This result suggests that high noise levels, which commonly occur in single-subject learning data (see, e.g., Heathcote, Brown, & Mewhort, 2000), may confound model selection based on goodness of fit. In particular, if measurement noise is high, a power function may provide a better fit than an exponential function can, even when the data are generated by an exponential function.

It is certainly true that high noise levels can favor more complex models over less complex models. Modification of simple goodness-of-fit measures to account for model complexity, and so facilitate unbiased model selection, is an area of intense research (e.g., Bozdogan, 2000; Grunwald, 2000; Myung, 2000; Zucchini, 2000). However, simple power and exponential functions have the same number of parameters and, therefore, similar complexity.[1] Hence, a difference in model complexity may not explain the very large bias found by Myung et al. (2000).

Myung et al. (2000) explained their observed bias using the "response surfaces" of the functions. The response surface, $S \subset R^M$, of an arbitrary function $f(t,\theta)$, where $t$ is $M$-vector of covariate values such as $t = \{1, 2, \ldots, M\}$ and $\theta$ is a $k$-vector of parameters, is defined as $\{x \in S \Leftrightarrow [x_i = f(t_i, \theta), \forall\, i = 1 \ldots M]\ \theta \notin R^k\}$. An observed curve, considered as a point in $R^M$, is "best fit" by the function whose response surface most closely approaches that point, usually in the sense of a Euclidean distance when ordinary least squares fitting is used. In Myung et al.'s analyses, the space of interest was $[0,1]^2$. Hence, they were able to explain the bias toward the power function in random data by noting that the power function's response surface is closer to the center of $[0,1]^2$, and so is closer to a larger volume of the space, than is the exponential function's response surface. As the binomial noise which Myung et al. examined fills this space symmetrically around its expected value (0.5, 0.5), the power function will, on average, provide a better fit.

We argue that Myung et al.'s (2000) result usually does not apply to empirical model selection issues, because of the restricted case that they examined. First, in order to make the response surface analysis (RSA) more tractable, Myung et al. looked at learning curves consisting of only two points, many fewer points than are usually measured empirically. Because RSA becomes intractable as the number of points increases, we use simulation methods to investigate more realistic cases. More importantly, Myung et al. examined only simple (one-parameter) power and exponential functions. Such functions are applicable only to situations where performance is measured by a dimensionless quantity, such as probability correct, and learning causes performance to change between two known levels. More commonly, either the initial or the final level of performance (or both) is unknown and so must be estimated from the data. Where the initial level of performance is unknown, a scale parameter ($B$) must be estimated ($y = B \cdot f(t) + \varepsilon$). Where both initial and final levels of performance are unknown, both location ($A$) and scale parameters must be estimated ($y = A + B \cdot f(t) + \varepsilon$). Examples of one-, two-, and three-parameter power and exponential functions are given in Figure 1. Note that the one-parameter functions (top row) start from the top of the $y$-axis and descend to zero. The two-parameter functions (middle row) start from lower than the top of the $y$-axis, but still

Correspondence concerning this article should be addressed to S. Brown, Department of Cognitive Sciences, University of California, Irvine, CA 92697-5100 (e-mail: scottb@uci.edu).

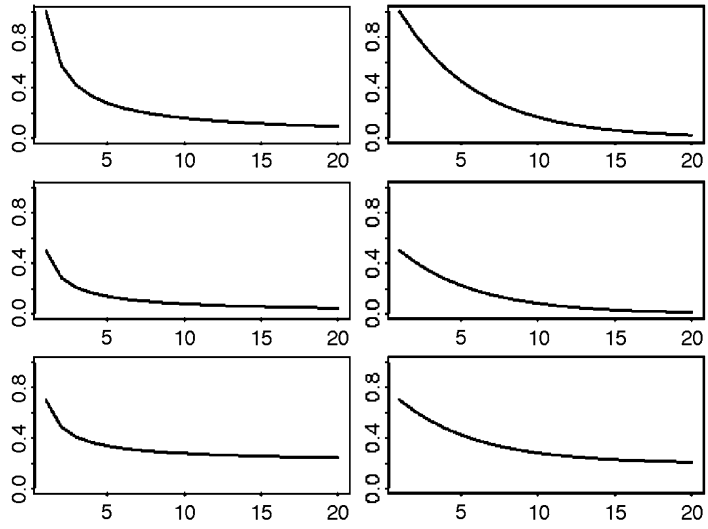


**Figure 1. Example power (left column) and exponential (right column) functions. The top row shows one-parameter functions ($y = x^{-0.8}$ and $y = e^{-0.2x}$), the middle row shows two-parameter functions ($y = 0.5x^{-0.8}$ and $y = 0.5e^{-0.2x}$), the bottom row shows three-parameter functions ($y = 0.2 + 0.5x^{-0.8}$ and $y = 0.2 + 0.5e^{-0.2x}$).**

descend to zero. The three-parameter functions (bottom row) also start from lower than the top of the *y*-axis and do not descend all the way to zero, no matter how far the *x*-axis is extended.

## SIMULATION 1
## Noise Only

In Simulation 1, we had three main aims. First, we replicated the work of Myung et al. (2000) by fitting one-parameter (simple) power and exponential functions to binomial noise. Second, we extended this result by considering the fits of two- and three-parameter power and exponential functions to binomial noise. Third, we repeated the simulations, using both normal and log normal noise. Binomial noise is of interest when the data are probabilities or percentage scores, as in memory retention studies. However, power and exponential function fits are also compared in other paradigms where binomial noise is an unreasonable assumption. We chose the log normal distribution because response time data from practice paradigms often have a skewed distribution, similar to the log normal, and power and exponential function comparisons have been of some importance in that paradigm (Heathcote et al., 2000; Newell & Rosenbloom, 1981). We also considered the normal distribution because of its limiting status as the sum of independent observations from other distributions. This would apply to paradigms where summed or averaged measures are used for fitting, such as "block averaged" response times, or mean percent correct measured over many items.

All simulations used scripts written by the authors in the S language (Becker, Chambers, & Wilks, 1988), implementing maximum likelihood estimation. We also repeated the calculations, using least squares estimation, with essentially the same results. Start points for the one-parameter functions were obtained by linear regression on either log transformed (for exponential functions) or log–log transformed (for power functions) data. Optimal parameter estimates from the one-parameter functions were used as degenerate (i.e., scale parameters were started from unity) start points for the two-parameter functions. Similarly, the two-parameter fits were used as degenerate (asymptote parameters started at zero) start points for the three-parameter functions.

Binomial data were generated as in Myung et al. (2000), by taking taking the mean of 50 independent Bernoulli samples. Error samples from the normal and log normal distributions were generated by using the algorithms provided in the S language. The expected value of the binomial data was 0.5, and its expected variance was 0.005. The parameters of the normal and log normal were chosen to match these values: The normal distribution thus had mean 0.5 and variance 0.005; the log normal distribution had the same mean and variance (based on an underlying normal distribution with mean −0.7030 and standard deviation 0.1407). In all simulations, the results from these three different error distributions were almost identical. We report only the binomial results for brevity.

For each error model, the best-fitting one- (simple), two- (scaled), and three- (located and scaled) parameter power and exponential functions were estimated. This process was repeated 1,000 times for each error model, and for each series length in the set {2, 3, 4, 5, 8, 15, 20, 30, 50, 75, 100}. From these fits, the proportion of curves best fit (i.e., with larger likelihood) by an exponential function was recorded. The two-parameter functions were not fit to curves of length two, and the three-

parameter functions were not fit to curves of length two or three. All simulation calculations were carried out at double precision (approximately 15 significant figures). Occasionally, the two functions being compared fit the data nearly equally well, with likelihood values equivalent to more than 7 significant figures. In such cases, we did not feel it fair to attribute a "winner" to either function, and so these were removed from further analysis. Such cases were infrequent (in total, less than 1.8% of the study).

Figure 2 summarizes the most important simulation results from fits to the binomial random data. The numbers on each line represent the number of parameters for the curves being compared (1 = simple, 2 = scaled, or 3 = located and scaled), the abscissa represents the length of the data series, and the ordinate represents the percentage of data series best fit by the exponential function.

The curve representing comparisons between the one-parameter models exhibits a strong bias toward the power function, a bias that increases with series length. This replicates and extends the bias found by Myung et al. (2000). However, the two- and three-parameter curves demonstrate little or no bias: Binomial random data series are about equally well described by these power and exponential functions. Thus, Myung et al.'s power function bias in noisy data seems to apply only to one-parameter power and exponential models.

Why should results analogous to Myung et al.'s (2000) fail to apply to exponential and power functions with more than one parameter? An approximate explanation, which provides some insight into this problem, is as follows. Given a three-dimensional response space $[0,1]^3$, the height of the response surface for a two-parameter exponential curve [i.e., $(x = f(t_1), y = f(t_2), z = f(t_3)$, where $f$ is the two-parameter exponential function] is given by $z = x^a y^{1-a}$, where $a = (t_2 - t_3)/(t_2 - t_1)$. The response surface for the corresponding two-parameter power function is given by exactly the same function, except with $a = (\ln t_2 - \ln t_3)/(\ln t_2 - \ln t_1)$. Given choices of $t_1$, $t_2$, and $t_3$ that respect $0 < t_1 < t_2 < t_3$, both these response surfaces intersect along the plane $x = y$, which includes the point (½, ½, ½), precisely the expected position of the binomial noise. Thus, both functions (power and exponential) provide on the average an equally good fit to binomial noise.

Attempts to extend the RSA heuristics provided above to the three-parameter curve (thus working in at least a four-dimensional space) are hampered by increasingly unmanageable algebra. However, because fitting a three-parameter power or exponential function to independent and identically distributed noise is a special case, we can provide a heuristic explanation as follows. Without attempting to determine the form of the response surface in a response space of arbitrary dimension greater than three, we can still be sure that the point in response space representing the expected value of the noise will lie on that surface, because the elements of the vector representing such a point are all equal (as the noise was assumed independent and identically distributed) and two- and three-parameter power and exponential curves can both fit precisely any set of constant values. That the expected value of the noise lies on the response surfaces

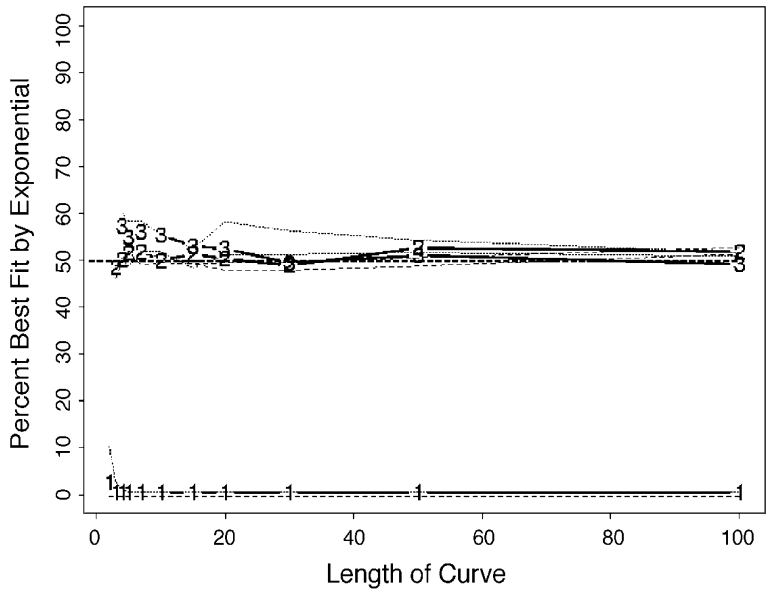


**Figure 2. The percentage of purely random data sets better described by exponential than by power functions. The abscissa represents the length of the data series. The numbers on each line correspond to the numbers of parameters for the functions being examined. The one-parameter functions compared were $y = x^{-\beta}$ and $y = e^{-\alpha x}$, the two-parameter functions compared were $y = Bx^{-C}$ and $y = Be^{-Dx}$, and the three-parameter functions compared were $y = A + Bx^{-\beta}$ and $y = A + Be^{-\alpha x}$.**

for both functions does not constitute a proof that they should fit equally well, but it does at least provide an aid to understanding why, all other things being equal, no bias would be expected.

The three-parameter curve in Figure 2 tells a similar story to that of the two-parameter curve: a nearly 50/50 split of best-fitting power and exponential curves. However, there was a slight bias toward the exponential for some curve lengths. This bias was small (e.g., never greater than 56% of curves identified as exponential) but was statistically reliable for curve lengths up to 10, owing to the large number of replicates in the simulations. This bias is most likely due to numerical difficulties. Because the data contained no structure, parameter estimates were underconstrained, resulting in poor search convergence behavior, especially for the power function. Informal observation supported this explanation. For example, the power function estimates on the average took much longer to converge to their minima, needing nearly twice as many iterations, even though the starting values were of similar quality. This bias also disappeared when cases in which the power and exponential fits were equal to single precision levels (around six significant figures) were removed from the analyses.

The results of our simulations suggest that a necessary condition for a bias to be avoided is the presence of at least one linear parameter in the functions being evaluated. Recall that a linear parameter in a given function is any parameter such that the derivative of the function with respect to that parameter is independent of the parameter. The two-parameter and three-parameter functions examined contain linear parameters, whereas the one-parameter functions do not. However, any function with linear parameters can be re-parameterized in such a way as to make all its parameters nonlinear. Such a re-parameterization does not change the function itself: The response surface of the function is unaltered by re-parameterization. Given the equivalence of such representations, it seems that a necessary condition to avoid biases is that the function in question have at least one linear parameter in at least one particular parameterization. This conclusion was supported by repeating the simulations above with re-parameterized versions of their functions. These re-parameterizations[2] were designed so that no function had any linear parameters. The three-parameter power function was reexpressed as $e^A + B[(x+1)/B]^{-C}$, and the corresponding exponential function as $e^A + e^{(B-Cx)}$. In both cases, the corresponding two-parameter functions were formed by omitting the first term, and the one-parameter functions were formed from those by setting $B = 1$ for the power function and $B = 0$ for the exponential. The results of these simulations confirmed those above: strong bias toward the power function for comparisons between the one-parameter functions, and essentially no bias for comparisons between two- or three-parameter functions.

The conclusion to be drawn from these analyses is that the biases observed above will not occur if there exists *any* re-parameterization in which the functions under examination have linear parameters. Note that some "re-parameterizations" of the functions above can exclude certain degenerate cases, and so may not lead to the same results as they effectively examine different response surfaces. For example, an apparent re-parameterization of the two-parameter power function $[(x+1)/B]^{-C}$ cannot predict the crucial expected data pattern of {½, ½, . . . , ½}. In light of the RSA analysis above, such a limitation would be expected to reduce the ability of the power function to fit random binomial data. Replication of our simulations using this parameterization for the power function confirmed the above hypothesis: There was a strong bias toward the *exponential* function for two-parameter comparisons.

## SIMULATION 2
## Signal Plus Noise

The "data" in the first set of simulations were entirely without structure: purely random noise. In practice, parametric functions are fit to data that contain both structure and noise. The second set of simulations extended the first by examining that situation. In particular, we aimed to determine whether the preceding "noise-only" preference rates emerge gradually as the ratio of structure to noise decreases. To generate the data for Simulation 2, we used both exponential (corresponding to dashed lines in Figure 3) and power (solid lines) functions to calculate probabilities for each point on the curve. These probabilities (say $y$) were then attenuated by weighting toward their mean value (0.5) according to $y \rightarrow y(1-w) + 0.5w$. Hence $0 < w < 1$ governs the level of attenuation applied to the probabilities. The resulting attenuated probabilities were used to generate 50 binomial samples for each point on the curve, which were then averaged as in Simulation 1. This algorithm was intended to model a situation in which an experimenter uses a (50-sample) binomial design to measure an underlying process that is subject to weakening or attenuation during the measurement process. We used five values of $w$ in Simulation 2: 0.0, 0.125, 0.25, 0.5, and 1.0; in Figure 3, these are labeled "a"–"e" respectively. Similar results (not reported here) were obtained with a second algorithm in which the expected probabilities ($y$) were perturbed according to $y \rightarrow y(1-w) + 0.5u$, where $u$ is a random deviate distributed uniformly in [0,1]. This second algorithm was intended to model a situation in which the experimenter's expected values were perturbed by random measurement noise.

The exponential and power functions used to calculate the underlying probabilities had rate parameters of 0.03 and 0.4, respectively. These values were chosen both to be similar to those observed in data (e.g., Rubin, Hinton, & Wenzel, 1999) and to provide a close match between exponential and power function expected values. For two- and three-parameter functions, the same location (0.4) and scale (0.2) parameters were used for both power and exponential functions.

Figure 3 shows the results for discriminations between one-parameter functions, for different noise levels

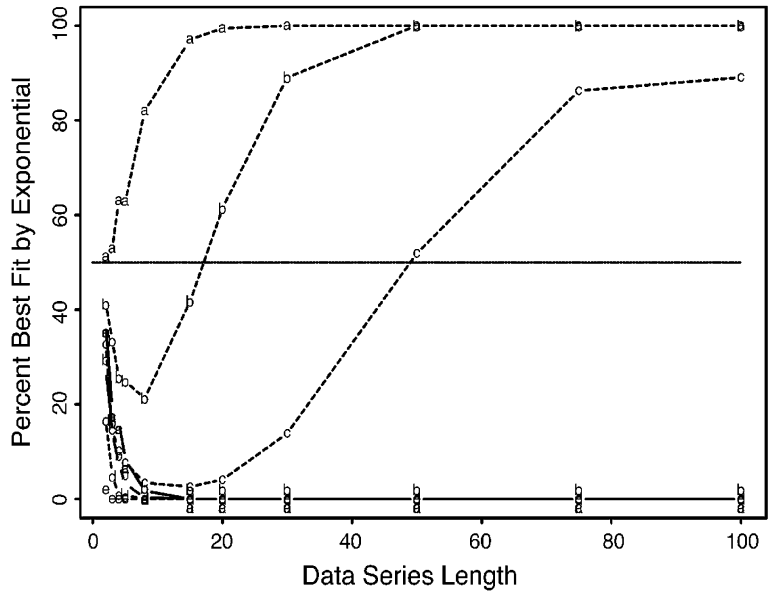


**Figure 3. Percentage of data sets containing both noise and structure better described by one-parameter exponential ( $y = e^{-Dx}$ ) than power ( $y = x^{-C}$ ) functions. Data were generated by one-parameter exponential (dashed lines) or power (solid lines) functions. Noise levels were 0.02 ("a"), 0.06 ("b"), 0.15 ("c"), 0.3 ("d"), and 0.5 ("e").**

("a"–"e") and underlying function type. In low noise conditions, power and exponential functions were reliably distinguished for most curve lengths. However, increased levels of noise produced results similar to those in Simulation 1. Note that power functions were better discriminated in low noise conditions, with power function discriminations being further from chance than exponentials for the same curve length. For example, in the condition "a," there was 99% accurate discrimination of power function data for all curves containing 8 or more observations, whereas exponential functions required 30 or more observations to reach this level. Even data from noise levels that were perfectly discriminable for longer curve lengths ("a"–"c") showed some bias at shorter lengths, where the majority of exponential functions were misidentified as power functions.

Similar simulations were carried out with two-parameter and three-parameter functions. The results for the two-parameter functions appear in Figure 4; those for three-parameter functions were almost identical and so are not graphed. Figure 4 shows a result analogous to that in Figure 3: at low levels of noise, there is reliable identification of curve form, but higher levels lead to 50/50 identification for two- and three-parameter functions, as was observed in Simulation 1.

Figures 3 and 4 illustrate a general trend for longer data sets to provide better discrimination, presumably because of an improved signal to noise ratio. However, in Figure 4, for higher noise levels ("c"–"e"), exponential data show better discrimination at medium curve lengths than at longer lengths. This difference may be explained through consideration of changes in the signal to noise ratio with increasing curve length. Exponential functions approach their asymptotic value much more quickly than power functions. Thus, for lengths beyond about 30 trials, extra points can contribute more noise but no new information about shape, and so can only reduce discriminability. The power function, in contrast, approaches its asymptote more slowly, so that longer series continue to introduce enough new shape information to counteract the extra noise. Even for the power function, however, a sufficiently long series would be expected to decrease discrimination eventually.

## CONCLUSIONS

Our results show that power and exponential functions with the same number of parameters *do* fit pure noise equally, at least where the functions have scale, or scale and location parameters, regardless of parameterization. The surprising and important bias toward the power function described by Myung et al. (2000) was replicated for the special case of one-parameter functions, and it appears to be very strong, especially for longer curves. The increasing bias with increasing curve length for one-parameter functions also fits naturally into the RSA explanation: The proportion of space occupied by the exponential curve would be expected to decrease geometrically as the dimension of the space (curve length) increases, and so the magnitude of the bias increases with curve length.

The same results were found to hold for random data generated from the binomial, normal, and log normal distributions, suggesting that our results may have im-

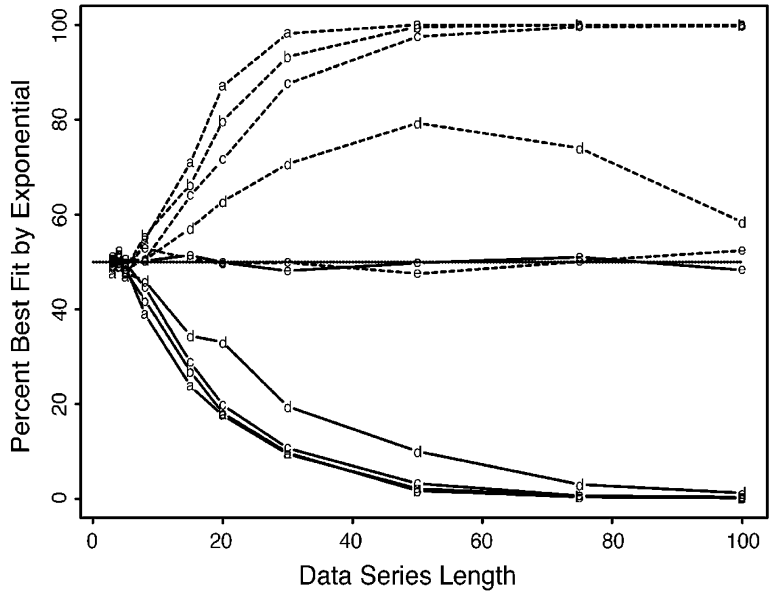


**Figure 4. Percentage of data sets containing both noise and structure better described by two-parameter exponential ($y = Be^{-Dx}$) than by power ($y = Bx^{-C}$) functions. Data were generated by two-parameter exponential (dashed lines) or power (solid lines) functions. Noise levels were 0.02 ("a"), 0.06 ("b"), 0.15 ("c"), 0.3 ("d"), and 0.5 ("e").**

portance for empirical model selection in paradigms where these distributions are typical. Importantly for empirical work in these paradigms, our results support the validity of model selection in noisy data based on comparative goodness of fit for two- and three-parameter exponential and power functions.

Our results also demonstrate that when power and exponential functions are present in data, but are obscured by noise, similar effects occur. In particular, high levels of noise lead to equal preference for power and exponential functions for comparisons between two- or three-parameter functions. For comparisons between one-parameter power and exponential functions, increasing levels of noise lead to strong biases toward power functions. This is a situation analogous to that of Myung et al. (2000), and one that researchers must be careful of.

As a corollary, our results imply that methods that reduce noise levels, such as averaging, will act as expected: improving model discrimination by pushing results away from equal preference, for two- and three-parameter functions, or reducing power bias, for simple functions. Note, however, that averaging across curves from different participants or different conditions can introduce a strong bias favoring the power over the exponential function. Averaging across contiguous trials within a single curve (block averaging), in contrast, produces little or no bias and so can be more safely used to improve model selection by decreasing noise (Brown & Heathcote, 2003).

## REFERENCES


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

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology*, **44**, 62-91.

Brown, S., & Heathcote, A. J. (2003). Averaging learning curves across and within participants. *Behavior Research Methods, Instruments, & Computers*, **35**, 11-21.

Grunwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, **44**, 133-152.

Heathcote, A. [J.], 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.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, **44**, 190-204.

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.

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.

Rubin, D. C., Hinton, S., & Wenzel, A. (1999). The precise time course of retention. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **25**, 1161-1176.

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, **44**, 41-61.


## NOTES

1. Complexity is not completely determined by the number of parameters to be estimated, as implied by simple complexity corrections such as AIC and BIC.

2. We are indebted to In Jae Myung for bringing this to our attention.

(Manuscript received January 22, 2002;
revision accepted for publication February 14, 2003.)