Behavior Research Methods, Instruments, & Computers
2003, 35 (4), 493-503

Testing curvatures of learning functions on individual trial and block average data

DENIS COUSINEAU, SÉBASTIEN HÉLIE, and CHRISTINE LEFEBVRE
Université de Montréal, Montréal, Québec, Canada

Many models offer different explanations of learning processes, some of them predicting equal learning rates between conditions. The simplest method by which to assess this equality is to evaluate the curvature parameter for each condition, followed by a statistical test. However, this approach is highly dependent on the fitting procedure, which may come with built-in biases difficult to identify. Averaging the data per block of training would help reduce the noise present in the trial data, but averaging introduces a severe distortion on the curve, which can no longer be fitted by the original function. In this article, we first demonstrate what is the distortion resulting from block averaging. The *block average* learning function, once known, can be used to extract parameters when the performance is averaged over blocks or sessions. The use of averages eliminates an important part of the noise present in the data and allows good recovery of the learning curve parameters. Equality of curvatures can be tested with a test of linear hypothesis. This method can be performed on trial data or block average data, but it is more powerful with block average data.

Many experiments involve training in a task. This is commonly done to reduce the variability that would arise from unskilled subjects. In this case, the experimenter is interested only in the final level of performance, often described by one or a few summary values (mean response time, standard deviation, percent correct, etc.). However, some researchers are not interested simply in a snapshot, but in the whole dynamic of performance over training (e.g., Logan, 1988; Rickard, 1997; Shiffrin & Schneider, 1977). Because of the large number of data involved, it is often convenient to summarize them in a curve: the learning curve (Heathcote, Brown, & Mewhort, 2000; Newell & Rosenbloom, 1981).

Learning curves describe the evolution of performance over trials t. They are given by the following equation:

$$f(t) = a + b \cdot g(t), \tag{1}$$

where a is the asymptote of the curve and b is the amplitude. These two scaling parameters act as boundaries, since initial performance is given by the value $a + b$ and final performance is given by a.[1] The function $g(t)$ describes the type of curvature present in the learning curve. As such, $g(t)$ is called the *core* of the learning curve and is often a function of a third parameter, the learning rate parameter c (Paul, 1994).

The purpose of this article is not to decide which type of learning curve best describes the data. This issue is still highly controversial. When performance has been measured by response times, many authors have defended the power curve (Logan, 1988; Newell & Rosenbloom, 1981). Its core function is given by $g_{PC}(t) = t^{-c}$. But Heathcote and his colleagues have raised some concerns over recent years (Heathcote et al., 2000). They have suggested that the exponential curve, given by the core $g_{EX}(t) = e^{-c\,t}$, was as good a contender. Other learning curves have also been proposed over the years, such as the general power curve [$g_{GP}(t) = (t + d)^{-c}$; Newell & Rosenbloom, 1981], which has a free parameter d to take into account learning prior to the beginning of the task (see also Cousineau, Goodman, & Shiffrin, 2002). In the context of memory research, the retention curve measuring percentage recalled as a function of time is also a function that fits the framework of Equation 1 (Wixted, 1990). Which core function is the correct one is not an issue that has been resolved. In addition to the theoretical question of which type of function describes the core, there is an empirical question about the curvature present in the performance, given a hypothesized core function. Curvature (or learning rate) is a measure of the speed at which performance reaches the asymptote. In the following, the curvature is quantified by the learning rate parameter c, assuming one type of curve (exponential, power, etc.).

Some theories predict that the stimuli to be learned will affect the curvature (reduction of information theories; e.g., see Haider & Frensch, 1996), whereas other theories predict that the stimuli will not affect curvatures

Part of this research was presented at the 11th Annual Meeting of the Brain, Behaviour and Cognitive Science Society, Québec (2001). This research was supported by the Fonds pour la formation de chercheurs et l'aide à la recherche, Établissement de nouveaux chercheurs. We thank Dominic Charbonneau, Andrew Heathcote, Guy Lacroix, Serge Larochelle, and an anonymous reviewer for their comments on an earlier version of this text. Correspondence concerning this article should be addressed to D. Cousineau, Département de psychologie, Université de Montréal, C. P. 6128, succ. Centre-ville, Montréal, PQ, H3C 3J7 Canada (e-mail: denis.cousineau@umontreal.ca).

but only the bounds a and b (such as strength theories; see Dumais, 1979). Logan's (1988) instance-based theory predicts that curvatures will be equal for the mean response times and their standard deviations. This same prediction also holds for the SSTS*, a subset of the serial self-terminating class of models in visual search (Cousineau & Larochelle, 2003). The aim of this article is twofold. First, because mean performances are often used, we present a simple method for recovering the parameters $\{a, b, c\}$ out of averaged performance. Second, we present a method by which to test whether the learning rates of two or more curves are equal. This method is applicable as soon as one type of core function is assumed. The core function can be any function that fits Equation 1 and so avoids the above controversy.

Fitting Averages

Most theories of learning assume that learning occurs on a trial-by-trial basis. Yet raw data (called *trial data* hereafter) are usually very erratic, making the learning curve hard to see. To reduce the noise present in the data, researchers usually aggregate their data over blocks of trials, using averages. However, Rickard (1997) has pointed out that the curve of the block averages generally does not have the same core as the curve of the trial data (as will be shown below). Yet this fact should not discourage the use of averaged data; we will show in this section how to obtain the learning function of a curve averaged over blocks of successive trials. As will be shown, fitting a curve of averaged data is as easy as fitting trial data but allows the recovery of the right parameters more efficiently.

Averaging curves. In what follows, we define $f(t)$ as the trial learning curve function, which is a function of the trial number t, going from 1 to T. We want to know what is the learning curve equation when the data are averaged over blocks of training. Let us define $\overline{f}(n)$ as the block average function over block number n when trial data are averaged in blocks of N trials each ($N > 0$ is a constant). Thus, n goes from 1 to T/N (T is assumed to be a multiple of N). In order to simplify the problem, we first examine the core function of the block average curve. Let $\overline{g}(n)$ be the block average core function. By definition of the arithmetic mean, we have

$$\overline{g}(n) = \frac{1}{N} \sum_{i=(n-1)N+1}^{nN} g(i),$$

where $g(i)$ is the core function in Equation 1 and i indexes all the N trials in the nth block. This equation generally cannot be simplified in the discrete case, but if N is large, we can solve it by using a continuous approximation:

$$\overline{g}(n) \approx \frac{1}{N} \int_{(n-1)N}^{nN} g(x)dx. \tag{2}$$

Equation 2 can be solved for many learning curves, yielding the equation of the block average core function.

Because a simple linear transformation relates the trial function and the core trial function, and because averages are not altered by such transformations, we can simply add the scaling parameters around the block average core function to obtain the full block average function:

$$\overline{f}(n) = a + b\ \overline{g}(n).$$

Scale invariant curves. A first question to ask is, which functions remain of the same type after averaging? In other words, which functions are scale invariant? This will answer Rickard's (1997) point, noted at the beginning of this section. Two scale invariant functions are easily identified, the first one being trivial: the line $[f(t) = a + b\ t]$ and the exponential curve $[f(t) = a + b\ e^{-ct}]$.

The line is a degenerate curve, since it has no curvature parameter. Its core function is simply $g_{\mathrm{LN}}(t) = t$. The scaling parameter b represents the slope, whereas a represents the intercept. By solving Equation 2 on g_{LN}, using blocks of size N, we obtain $\overline{g}_{\mathrm{LN}}(n) = Nn - N/2$. Thus, $\overline{f}(n) = a + b(Nn - N/2) = (a - bN/2) + (bN)n$. By substituting $a - bN/2 \rightarrow a'$ and $bN \rightarrow b'$, we obtain $\overline{f}(n) = a' + b'n$ and see that the block average core function is of the same type as the trial core function. One difference is that the slope is now steeper because it is expressed in different units (blocks vs. trials).

Similarly, we show that the exponential curve is also scale invariant. Solving Equation 2 on g_{EX}, we find that its block average core function is given by

$$\overline{g}_{\mathrm{EX}}(n) = \frac{e^{-cN(n-1)} - e^{-cNn}}{c\,N}.$$

Factorizing the exponential to isolate the dependent variable n, we obtain

$$\overline{g}_{\mathrm{EX}}(n) = \frac{\left(e^{cN} - 1\right)}{cN} e^{-cNn}. \tag{3}$$

By substituting

$$\frac{\left(e^{cN} - 1\right)}{cN} \rightarrow b'$$

and $c\,N \rightarrow c'$, we have $\overline{g}_{\mathrm{EX}}(n) = b'e^{-c'n}$. Thus, we see that the block average function of an exponential curve is also an exponential curve. With the scaling parameters a and b, this is a three-parameter curve $\{a, b, c\}$ for a given block size N.

Scale-dependent curves. The famous power curve is scale dependent since, as will be shown below, the core function is not functionally the same as the block average function. The core of the power function is given by $g_{\mathrm{PC}}(t) = t^{-c}$. Averaging the function over blocks of size N, using Equation 2, we obtain

$$\overline{g}_{\mathrm{PC}}(n) = \frac{\left[N(n-1)\right]^{-(c-1)} - (Nn)^{-(c-1)}}{(c-1)N}.$$

Note that $N(n-1)$ is the first trial of the nth block and Nn is the last trial of that block.[2] We therefore substitute $N(n-1) \rightarrow n_F$ and $Nn \rightarrow n_L$ to obtain

$$\overline{g}_{PC}(n) = \frac{n_F^{-(c-1)} - n_L^{-(c-1)}}{(c-1)N}. \quad (4)$$

Adding scaling parameters a and b, as in Equation 1, we see that $\overline{f}_{PC}(n)$ is a three-parameter curve defined by $\{a, b, c\}$, given a certain block size N. Therefore, it can be fitted to averaged data, using n_F and n_L, instead of the block number n, with no more difficulty than fitting a power curve.

Equation 4 is a difference between two power curves (or more precisely, the same power curve at two different moments). Yet the core is functionally different from a power curve's core function $[\overline{g}(x) \uparrow g(x)]$. Thus, fitting block average data with the trial function should result in (1) poor fit and (2) noninterpretable learning rate parameters.

As an example, in the top part of Figure 1, we generated simulated response times (SRT) with a power curve over 400 trials, using the parameters $\{a = 0, b = 350, c = 0.455\}$. As was expected, a power curve fits the trial data perfectly, and a minimization algorithm (such as PASTIS; Cousineau & Larochelle, 1997) can recover the parameters almost perfectly (with a precision of ±0.1%). In the bottom part of Figure 1, the SRT were averaged into 10 blocks of $N = 40$ trials. The dotted line shows the best-fitting power curve. As can be seen, the power curve

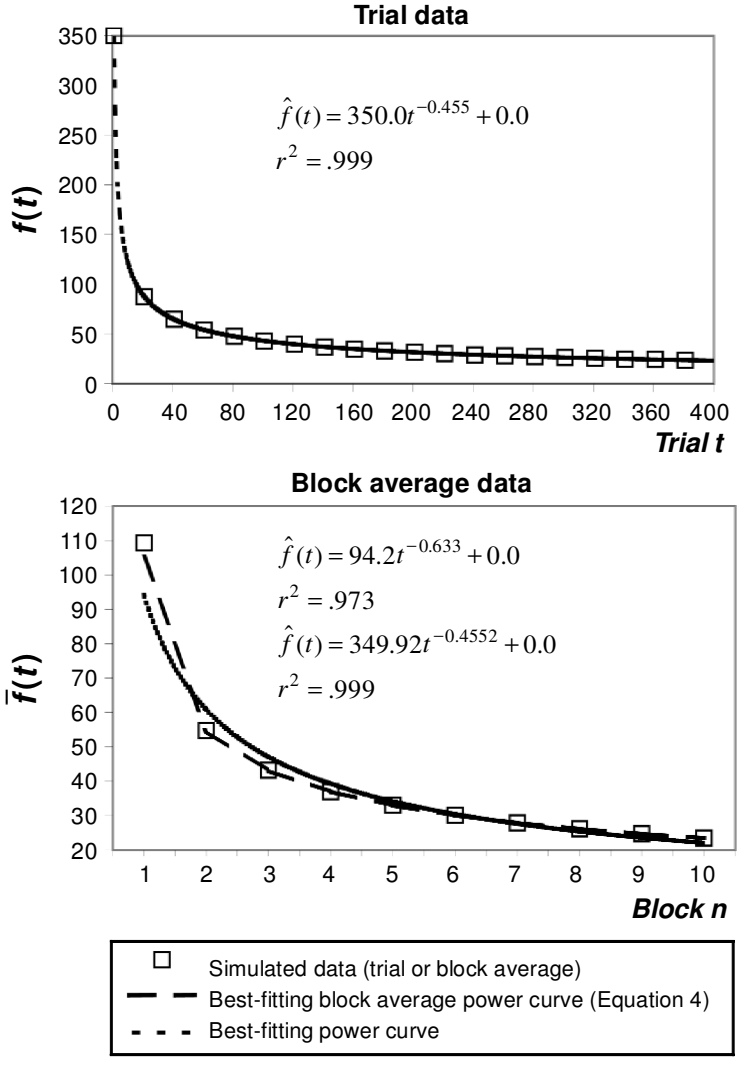

Figure 1. Averaging power curve per block. The top part shows a power curve generated using the parameter $\{a = 0, b = 350, c = 0.455\}$ over 400 trials (only a few simulated response times are shown with open boxes). The bottom part shows the same curve when averaged by blocks of $N = 40$ trials.

shows systematic deviations (poor fit, considering that there is no noise; $r = .973$), and the estimated parameters $\{\hat{a} = 0.00, \hat{b} = 94.2, \hat{c} = 0.633\}$ bear no resemblance to the true parameters. The dashed line shows the best-fitting block average power curve (Equation 4). The fit is almost perfect, even though we introduced a continuous approximation. Furthermore, the minimization algorithm recovered the parameters with a precision of ±0.1%. This shows that in the absence of noise, fitting the block average curve on averaged data is not more difficult than fitting the simpler trial data curve on trial data.

The first part of Appendix A explores the efficiency with which the block average function recovers the parameters when noise is present. It shows that, in general (trial data or block average data), the major factor that makes parameters difficult to recover is noise. The impact of noise can be reduced significantly by increasing the number of trials. The second part of Appendix A shows that it is preferable to use block averages when fitting parameters if the curvature is steep (c is bigger than 0.4).

Illustrating the core function. One convenient way to look at curvature is to have a graph of the core function. Remember that all core functions start at one and have an asymptote of zero. Thus, if the curves have an equal learning rate, their core functions should superimpose. Furthermore, if block average data are plotted, standard error (*SE*) intervals around the core functions can be computed.

In order to plot the core function, one must first choose which curve is assumed to underlie the data. For example, it can be the power function, in the case of trial data, or Equation 4 if block average data are used. Isolating the core of a learning curve requires that each point at time x (trial or block number) be transformed using

$$\hat{g}(x) = \frac{f(x) - \hat{a}}{\hat{b}}, \tag{5}$$

where $\hat{a}$ and $\hat{b}$ are estimates of the two scaling parameters $\{a, b\}$ and f is the observed performance at time x. If both $\hat{a}$ and $\hat{b}$ are valid estimators, Equation 5 returns a valid approximation of the core function $\hat{g}$.

If summary values are plotted (such as mean or standard deviation), the *SE* intervals can be computed (this approach cannot be used with trial data). *SE* can be used as a general indicator of whether two curves superimpose or not:

SE of the block average data at block n is given by

$$SE_{\bar{f}}(n) = \frac{\overleftrightarrow{f}(n)}{\sqrt{N}},$$

where N is the number of observations per block and $\overleftrightarrow{f}(n)$ is the estimated standard deviation at block n (Cramér, 1946). Equation 5 requires *SE* for transformed scores, but manipulating *SE* is well established (Tremblay & Chassé, 1970). For example, adding a constant to a score does not alter its *SE* interval, whereas multiplying it by a constant multiplies its *SE* interval. The estimated block average core function is thus given by

$$\hat{\bar{g}}(n) = \frac{\bar{f}(n) - \hat{a}}{\hat{b}} \pm \frac{\overleftrightarrow{f}(n)}{\hat{b}\sqrt{N}},$$

where $\bar{f}(n)$ and $\overleftrightarrow{f}(n)$ are the average and the standard deviation of the empirical measures at block n. Equivalent manipulations can be performed for any summary value normalized according to Equation 5, as long as its *SE* is known (Kendall & Stuart, 1983).

Illustrating the core function might provide an interesting solution to the related question, Did performance reach the asymptote? Formally, the performance will never reach asymptote since, for most learning curves, it requires an infinite amount of practice. Nevertheless, subjects may reach a level at which performance does not significantly differ from asymptotic performance. A very stringent criterion could be to declare a priori that asymptotic performances are reached if the core function is within 2 *SE*s of zero on the last four blocks.

Testing Curvatures

In this section, we will describe a method for testing whether two or more curvatures are equal, irrespective of the scaling parameters (amplitude and asymptote). Consider the following curves: $f_1, f_2, \ldots, f_s$ with unknown parameters $\{a_i, b_i, c_i\}$ for the ith curve. The most intuitive method for testing whether the curvatures are equal would consist in estimating the curvatures $\hat{c}_i$ (using a minimization procedure) and comparing them with a statistical test. However, this method has a very low power, because a lot of information is lost (a large data set is compressed into a single estimate $\hat{c}_i$). Considering that, in general, experiments involving learning have only a few subjects, this compression is too important.

The test of linear hypothesis (Rao, 1959) avoids this problem because it constrains the fit on more than singleton c_i. Suppose that s data sets are available, forming s learning curves labeled f_1 to f_s. If the core functions g_i are all identical, we can write

$$\begin{aligned} f_1(t) &= a_1 + b_1 g(t) \\ f_2(t) &= a_2 + b_2 g(t) \\ &\cdots \\ f_s(t) &= a_s + b_s g(t). \end{aligned}$$

As a consequence, we can show that the average curve $f_{\bar{s}}$ is given by the average parameters and the core function:

$$f_{\bar{s}}(t) = E(a_i) + E(b_i) g(t),$$

where $E(a_i)$ is the average of the a_i and $E(b_i)$ is the average of the b_i, $i = 1 \ldots s$. If the average curve $f_{\bar{s}}$ does not capture the data, it means that the core function is not unique to the s data sets. This is called a linear hypothesis.

One method by which to test whether the curve with averaged parameters captures the average data set is the linear hypothesis test created by Rao in 1959. It has been mentioned in Paul (1994), but with minimal details. One

objective of this section is to detail the structure of the test and to provide a short Mathematica listing that performs it (Wolfram, 1996). However, the real contribution of this section is to use the block average learning curve in conjunction with Rao's test and to show that doing so drastically increases the power of the test.

Applying the test of the linear hypothesis to trial data. Following Rao (1959), the first step is to describe the model underlying the data set. In terms of vectors, let the model be $\mathbf{M} = \{1, g(t)\}$ and the parameters $\theta = \{\alpha, \beta\}$, so that $\theta^T\mathbf{M} = \alpha + \beta\, g(t)$. It must be understood that $g(t)$ is also a function of c, the learning rate parameter. Suppose that we have collected, for each of the s data sets, a number T of trial observations. The model $\mathbf{M}$ varies according to the trial number. The matrix $\mathbf{A}$ summarizes the evolution of the model for each trial and each parameter. We can write

$$\mathbf{A} = \begin{pmatrix} 1 & g(1) \\ 1 & g(2) \\ & \cdots \\ 1 & g(T) \end{pmatrix},$$

where the first column indicates the contribution of α to the performance of the average curve and the second column indicates the contribution of β.

In the implementation of the model, c is not considered a parameter. Therefore, it must receive a value at this point. However, under the null hypothesis, every set has the same curvature, and the average curve is also representative of the curvature. Thus, a numerical value for c should be obtained, using a least-square minimization routine (such as PASTIS; Cousineau & Larochelle, 1997) on the between-sets average data.

The next step is to obtain the set of estimates $\hat{\boldsymbol{\theta}}$ that offers the best fit. Rao (1959) proposed one method,[3] but it is our experience that a better approach (less biased) is to take advantage of the null hypothesis that says that the group best-fitting parameters $\{\hat{\alpha}, \hat{\beta}\}$ ought to be the average of the individual subject best-fitting parameters. So let $\hat{\boldsymbol{\theta}} = \{\hat{\alpha} = E(\hat{a}_i), \hat{\beta} = E(\hat{b}_i)\}$. In summary, (1) fit the average curve to obtain the curvature c, and (2) fit the individual curve and average the individual asymptotes and amplitudes to obtain the parameter set $\hat{\boldsymbol{\theta}}$. The estimate $\hat{\boldsymbol{\theta}}$ is valid only if the null hypothesis is not rejected.

In order to perform a statistical test, summary values are needed. The first summary value is a vector, $\mathbf{y} = \{E[f_i(1)], \ldots, E[f_i(T)]\}$, containing the between-subjects average performance for the various trials from 1 to T. The second summary value is a variance–covariance matrix (of size $T \times T$), called hereafter $\mathbf{S}$, such that [*see equation at bottom of page*] where $\mathrm{Var}[f_i(j)]$ is the unbiased variance of the performances at time j and $\mathrm{Cov}[f_i(j), f_i(k)]$ is the unbiased covariance of the observations between trials at time j and trials at time k. This matrix is symmetrical.

The following equation is used to test the significance of the linear hypothesis. Let r be the number of data points in each of the curve T minus the number of parameters (generally three) and n the number of data set s. The test is of the form

Reject H_0 if:

$$\mathcal{F} = \frac{n-r}{r} \times \left(\mathbf{y} - \mathbf{A}\hat{\boldsymbol{\theta}}\right)\mathbf{S}^{-1}\left(\mathbf{y} - \mathbf{A}\hat{\boldsymbol{\theta}}\right) > F(\alpha, r, n-r),$$

where $F(\alpha, r, n-r)$, the critical value for the decision at level α—say, 5%—is read on a Fisher F table with r, $n-r$ degrees of freedom for the numerator and the denominator, respectively. In cases in which the inverse cannot be found ($\mathbf{S}$ is singular), a pseudo-inverse can be used (Rao, 1959).

Overall, Rao's (1959) test of linear hypothesis requires (1) the type of learning function to fit, (2) a minimization procedure for finding the group curvature and the individual asymptotes and amplitudes, (3) summary values (a vector of mean performance at trial $t \neq T$ and a $T \times T$ variance–covariance matrix), and (4) extensive matrix manipulation capabilities. This last point used to be the most difficult to obtain. Rao has described a complex method for making optimal use of the desk calculator available at that time (to the point that the article is difficult to decipher). Schneiderman and Kowalski (1985) have described an implementation of the test, using SAS. Yet this program is still difficult to follow. In Appendix B, we present a short Mathematica program for computing the summary values ($\mathbf{y}$ and $\mathbf{S}$), the best-fitting parameter $\hat{\boldsymbol{\theta}}$, and the statistic $\mathcal{F}$.

This approach is more powerful than the intuitive ones described at the beginning of the section, because it does not reduce the data to a single value (c or r^2). In fact, when the hypothesis is tested, all the points along the curves are used as constraints to see whether the instantiated model $\mathbf{A}$ is capturing the individual observations.

As can be seen from the degrees of freedom, the test of the linear hypothesis requires that the number of data sets (generally, the number of subjects) be at least equal to the number of trials. Because a typical experiment often involves hundreds of trials, the number of subjects rapidly becomes prohibitive. As will be shown next, collapsing the trial data into a fewer number of blocks allows one to measure a smaller number of subjects and still have a powerful test.

Applying the test of linear hypothesis to block average data. First, we note that after block averaging has

$$\mathbf{S} = \begin{pmatrix} \mathrm{Var}[f_i(1)] & \mathrm{Cov}[f_i(1), f_i(2)] & \cdots & \mathrm{Cov}[f_i(1), f_i(T)] \\ \mathrm{Cov}[f_i(2), f_i(1)] & \mathrm{Var}[f_i(2)] & & \\ \vdots & & \ddots & \\ \mathrm{Cov}[f_i(T), f_i(1)] & & & \mathrm{Var}[f_i(T)] \end{pmatrix},$$

been performed, the s data sets now form s curves $\overline{f}_i$. These block average curves are not of the same type as the trial curves (unless they are scale-invariant functions). However, their core functions $\overline{g}_i(n)$ are known (e.g., it is Equation 4 in the case of a power curve). As such, under the null hypothesis that the curvatures are the same, we can write

$$\overline{f}_1(n) = a_1 + b_1\,\overline{g}(n)$$
$$\overline{f}_2(n) = a_2 + b_2\,\overline{g}(n)$$
$$\cdots$$
$$\overline{f}_s(n) = a_s + b_s\,\overline{g}(n).$$

Here, a_i and b_i are exactly the same as with the trial data. Thus, if all the curves have the same curvature (same core), we can also write

$$\overline{f}_{\bar{s}}(n) = E(a_i) + E(b_i)\overline{g}(n),$$

where $\overline{f}_{\bar{s}}$ is the average across data sets of the block averages. Here, we have two distinct averagings: first, within data set, to obtain the block average curves, and next, between the block average curves, to obtain a single $\overline{f}_{\bar{s}}$ curve. Also note that the relation between the block average curves ($\overline{f}_{\bar{s}}$ vs. the various $\overline{f}_i$) is the same as the relation between the trial data curves ($f_{\bar{s}}$ vs. the various f_i), one of a linear relationship. Hence, the test of the linear hypothesis is relevant here for the same reasons as those for the trial data.

The model is $\overline{\mathbf{M}} = \{1, \overline{g}(n)\}$ with parameters $\theta = \{\alpha, \beta\}$ from which we can create the matrix $\overline{\mathbf{A}}$ instantiating the model.

As an example, if we assume that the trial data follow a power curve, the instantiation for block average data, following Equation 4, is composed of lines for each block n of the sort

$$\left\{1, \frac{n_F^{-(c-1)} - n_L^{-(c-1)}}{(c-1)N}\right\},$$

where c must be determined using least-square methods, N is the number of trials per blocks, n_F is the first trial of block n (given by $N \times (n-1)$), and n_L is the last trial of block n (given by $N \times n$). If there are T observations in the trial data sets, there are T/N blocks in the block average data sets. Thus, the final matrix $\overline{\mathbf{A}}$ could be

$$\overline{\mathbf{A}} = \begin{pmatrix} & \\ 1 & \dfrac{(0N)^{-(c-1)} - (1N)^{-(c-1)}}{(c-1)N} \\ & \\ 1 & \dfrac{(1N)^{-(c-1)} - (2N)^{-(c-1)}}{(c-1)N} \\ & \cdots \\ 1 & \dfrac{(T/N-1)^{-(c-1)} - (T/N)^{-(c-1)}}{(c-1)N} \end{pmatrix}.$$

The matrix $\overline{\mathbf{A}}$ may look quite cumbersome. Yet, given c and N, it is easy to compute. In addition, it is now N times shorter, speeding up the remaining computations by a factor of N.

Whether we fit the trial data with the trial function or the block average data, using the block average function, the best-fitting parameters $\hat{\theta}$ should be identical. However, reducing the number of points tested using blocks makes it possible to measure a reasonable number of subjects. This would suggest that having very few blocks containing a lot of trials each is desirable (so that few subjects are required). This is not true; there is a tradeoff between blocks of increasing size and power. At some point, the blocks are so large that there are only a few blocks left. A reasonable compromise is to choose a block size N near the square root of the total number of trials. In the third section of Appendix A, we test this claim with Monte Carlo simulations.

Discussion

The advantages of fitting average curves are numerous. The average data are less noisy than the trial data. It is therefore possible that the parameters $\{\hat{a}, \hat{b}, \hat{c}\}$, estimated from the average data, will be more accurate (as is shown in Appendix A). Furthermore, for many popular core functions, the block average function $\overline{f}(n)$ is not more complex or more difficult to fit using a minimization algorithm (and Equation 3 or 4). In particular, it has exactly the same number of free parameters. We updated the learning curve estimation program PASTIS to fit the block average functions (source code available at http://mapageweb.umontreal.ca/cousined/papers/02-pastis). However, it still requires that the modeler make an assumption about which type of curves (power, exponential, or other) underlies the data. Finally, when block average data are used, standard errors can be computed around the core function.

The form of averaging presented here is a within-subjects average. As has been shown by Estes (1956), between-subjects averaging is risky if the individual subjects have different learning rates c. Indeed, the average of $f_1, f_2, \ldots, f_s$ cannot be solved unless the individual cs are known or are all equal. In the second section, we presented a test of curvature based on Rao's (1959) test of the linear hypothesis, which can be used to decide whether the curvatures are equal or not.

REFERENCES

BATES, D. M., & WATTS, D. G. (1988). *Nonlinear regression analysis and its application*. New York: Wiley.

COUSINEAU, D., GOODMAN, V., & SHIFFRIN, R. M. (2002). Extending statistics of extremes to distributions varying on position and scale, and implication for race models. *Journal of Mathematical Psychology*, **46**, 431-454.

COUSINEAU, D., & LAROCHELLE, S. (1997). PASTIS: A program for curve and distribution analyses. *Behavior Research Methods, Instruments, & Computers*, **29**, 542-548.

COUSINEAU, D., & LAROCHELLE, S. (2003). *Visual-memory search: An interogative perspective*. Manuscript submitted for publication.

CRAMÉR, H. (1946). *Mathematical methods of statistics*. Princeton: Princeton University Press.

DUMAIS, S. T. (1979). *Perceptual learning in automatic detection: Processes and mechanisms*. Unpublished doctoral dissertation, Indiana University, Bloomington.

ESTES, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, **53**, 134-140.
HAIDER, H., & FRENSCH, P. A. (1996). The role of information reduction in skill acquisition. *Cognitive Psychology*, **30**, 304-337.
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.
HOEL, P. G. (1964). Methods for comparing growth type curves. *Biometrics*, **20**, 859-872.
KENDALL, M. G., & STUART, A. (1983). *The advanced theory of statistics*. London: Griffin.
LOGAN, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, **95**, 492-527.
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.
PAUL, L. M. (1994). Making interpretable forgetting comparisons: Explicit versus hidden assumptions. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 992-999.
RAO, C. R. (1959). Some problems involving linear hypotheses in multivariate analysis. *Biometrika*, **46**, 49-58.
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.
SCHNEIDERMAN, E. D., & KOWALSKI, C. J. (1985). Implementation of Rao's one-sample polynomial growth curve model using SAS. *American Journal of Physical Anthropology*, **67**, 323-333.
SHIFFRIN, R. M., & SCHNEIDER, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, **84**, 127-190.
TREMBLAY, L.-M., & CHASSÉ, Y. (1970). *Introduction à la méthode expérimentale*. Montreal: Centre Educatif et Culturel Inc.
WIXTED, J. T. (1990). Analyzing the empirical course of forgetting. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 927-935.
WOLFRAM, S. (1996). *The mathematica book* (3rd ed.). New York: Cambridge University Press.

NOTES

1. The point at which the initial performance is measured depends on the type of curve. For the exponential curve, it is measured at time $t = 0$, and for the power curve, at time $t = 1$.

2. Actually, $N(n-1)$ returns zero as the first trial of the first block. For the power curve, it is inappropriate since, according to this type of curve, the performance is infinite at time $t = 0$. To solve this issue, we used $N(n-1) + ½$ and $N\,n + ½$ when doing actual fitting. Thus, blocks range from ½ to $N + ½$, $N + ½$ to $2N + ½$, and so forth.

3. With the model implementation **A** and the summary values **y** and **S** (see below), we can obtain the optimal parameter $\hat{\boldsymbol{\theta}}$ for the group by solving $\hat{\boldsymbol{\theta}} = (\mathbf{A}^T\mathbf{S}^{-1}\mathbf{A})^{-1}\,\mathbf{A}^T\mathbf{S}^{-1}\mathbf{y}$, which yields the least mean square solution to the problem (Bates & Watts, 1988). This method is based on the postulate that the differences between subjects remains the same with practice. It is not the case, since between-subjects variability diminishes with training (Cousineau & Larochelle, 2003).

APPENDIX A
Fitting and Testing Curves Using Trial and Block Average Data

The general objectives of this article are to describe a method by which to estimate curvatures and test them. These objectives are crucially dependent on a minimization algorithm that reduces the sum of square error (*SSE*) between the data and the ideal curve passing through the points. The parameters $\hat{\boldsymbol{\theta}} = \{\hat{a}, \hat{b}, \hat{c}\}$ that minimize the *SSE* are termed the best-fitting parameters.

Simulation 1: Testing Biases Using Trial Data

To explore the capabilities of a minimization algorithm to estimate the true parameters θ, we ran Monte Carlo simulations. We used the minimization software PASTIS (Cousineau & Larochelle, 1997), but we also tested the minimization procedure *FindMinimum*, implemented in Mathematica, and found no differences in the patterns of results. We present the results by using the following measures of bias: the average Euclidian distance between the ith estimates $\hat{\boldsymbol{\theta}}$ and the true parameters θ, obtained over a large number of replications. Bias can also be seen as the distance between the center of gravity of all the $\hat{\boldsymbol{\theta}}_i$ and the true θ ($i = 1 \ldots R$, the number of replications):

$$\text{Bias} := \|E(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta})\| = \|E(\hat{\boldsymbol{\theta}}_i) - \boldsymbol{\theta}\|,$$

where $\|x - y\|$ denotes the Euclidian distance between x and y in a three-dimensional space. To express the bias as a percentage, we divided this value by $\|\theta\|$. In addition, we computed the efficiency, a measure of dispersion around the true parameters θ:

$$\text{Efficiency} := SD\left(\|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}\|\right) = \frac{1}{R-1}\sum_{i=1}^{R}\|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}\|^2.$$

We generated power curve trial data. We kept the true asymptote constant at $a = 300$ and the true amplitude at $b = 1{,}000$. Because these are linear parameters, they would provide little information if they were varied. However, we varied the learning rate, because curves with almost nonexistent curvatures might be more difficult to fit than curves with pronounced descent. We used $c = \{0.2, 0.4, 0.6, 0.8\}$. We also added a small amount of noise to the generated curves. We used normal additive noise with zero mean and standard deviation η times the height of the curve minus the asymptote. The values η used were $\{0.5, 1.0, 2.0\}$. A value of 2.0 represents a large variability that is similar to typical human response time data. At ∞, the curve would reach the asymptote (height of zero), and so noise would be zero, but of course, we never generated that many points. The number of points generated, T (sample size), was varied $\{50, 100, 200, 400, 800, 1{,}600\}$. Each point represents one trial, starting at Trial 1. Table A1, column 2 recapitulates the factors.

For a given combination of curvature × sample size × noise, we generated a noisy curve and ran a minimization algorithm (PASTIS) to obtain the best-fitting parameters. We replicated this a thousand times, after which bias and efficiency were computed.

The results are shown in Figure A1. As can be seen, noise had an important impact on bias and efficiency. The more noise, the less accurate were the best-fitting parameters. It was still a reasonably small bias on average, since a typical set of estimated parameters was rarely more than 2% inaccurate. Sample size also had an important impact. Larger sample sizes tended to be less biased. Finally, the learning rates (small vs. steep) had no influence on the best-fitting parameters.

APPENDIX A (Continued)

Table A1
Overview of the Monte Carlo Simulations Performed in Appendix A

Description	Simulation 1	Simulation 2	Simulation 3
Purpose	Is bias and efficiency dependent on noise, curvature?	Is bias and efficiency improved by block averages?	Is test of linear hypothesis more powerful with averaged data?
Dependent measures	bias	bias	Type I and Type II errors
Factors varied	curvature sample size noise	block sizes curvature noise	curvature of Curve 1 curvature of Curve 2 block sizes
Factors held constant		sample size (400)	sample size (400) noise (2.0)

Note—Curvature levels are 0.2, 0.4, 0.6, and 0.8. Sample sizes T are 50, 100, 200, 400, 800, and 1,600. Noise levels η are 0.5, 1.0, and 2.0. Block sizes N are 1 (no block average), 5, 10, 20, 40, and 80.

Simulation 2: Testing Biases Using Block Average Data

The above simulations were performed using trial data. Next, we wanted to see whether there would be an improvement in the best-fitting parameters when we used block average data instead of trial data. We ran a second series of simulations, for which we used both trial data and block average data. The size of a block, N, was 5, 10, 20, 40, or 80 trials per block. To keep the number of results manageable, we fixed the sample size at

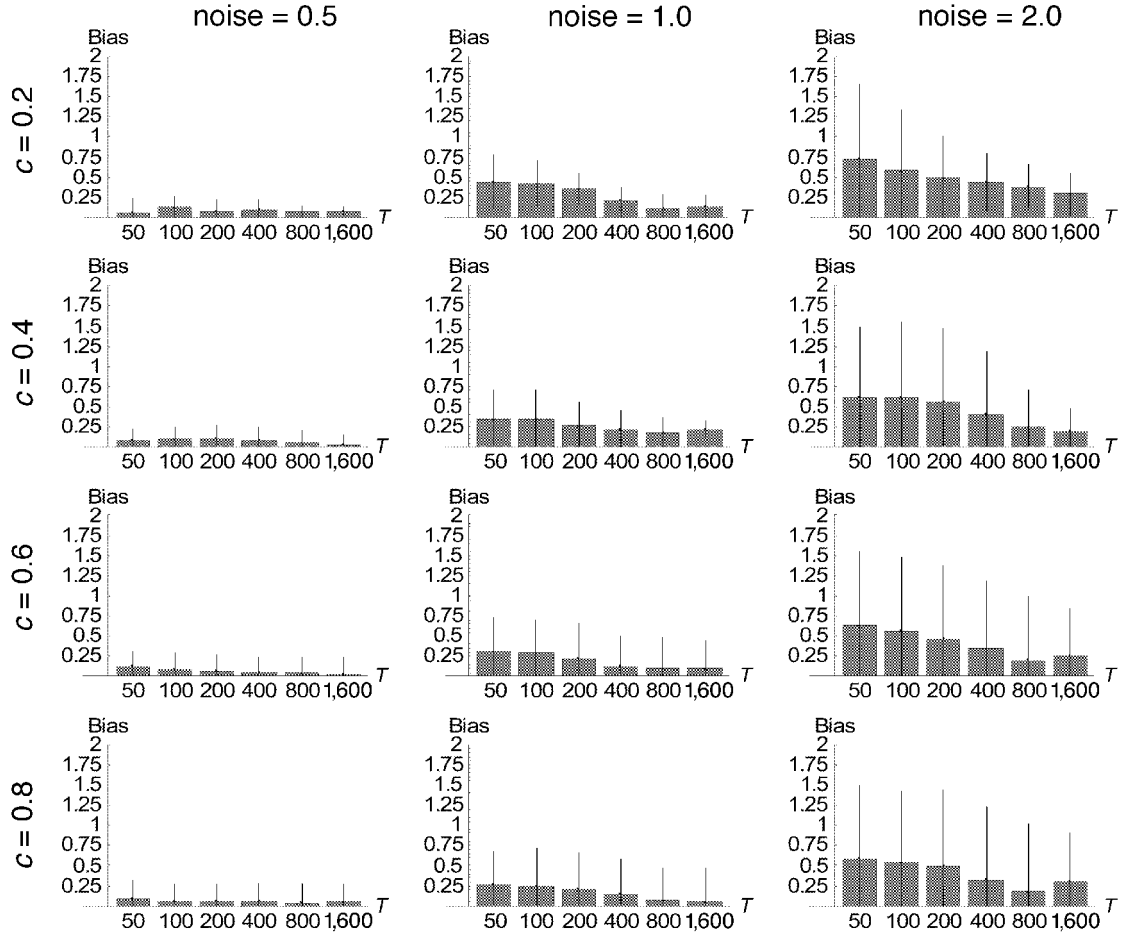

Figure A1. Bias and efficiency in percentages as a function of the number of trials T for curvature parameter c, increasing from top to bottom, and noise level η, increasing from left to right.

APPENDIX A (Continued)

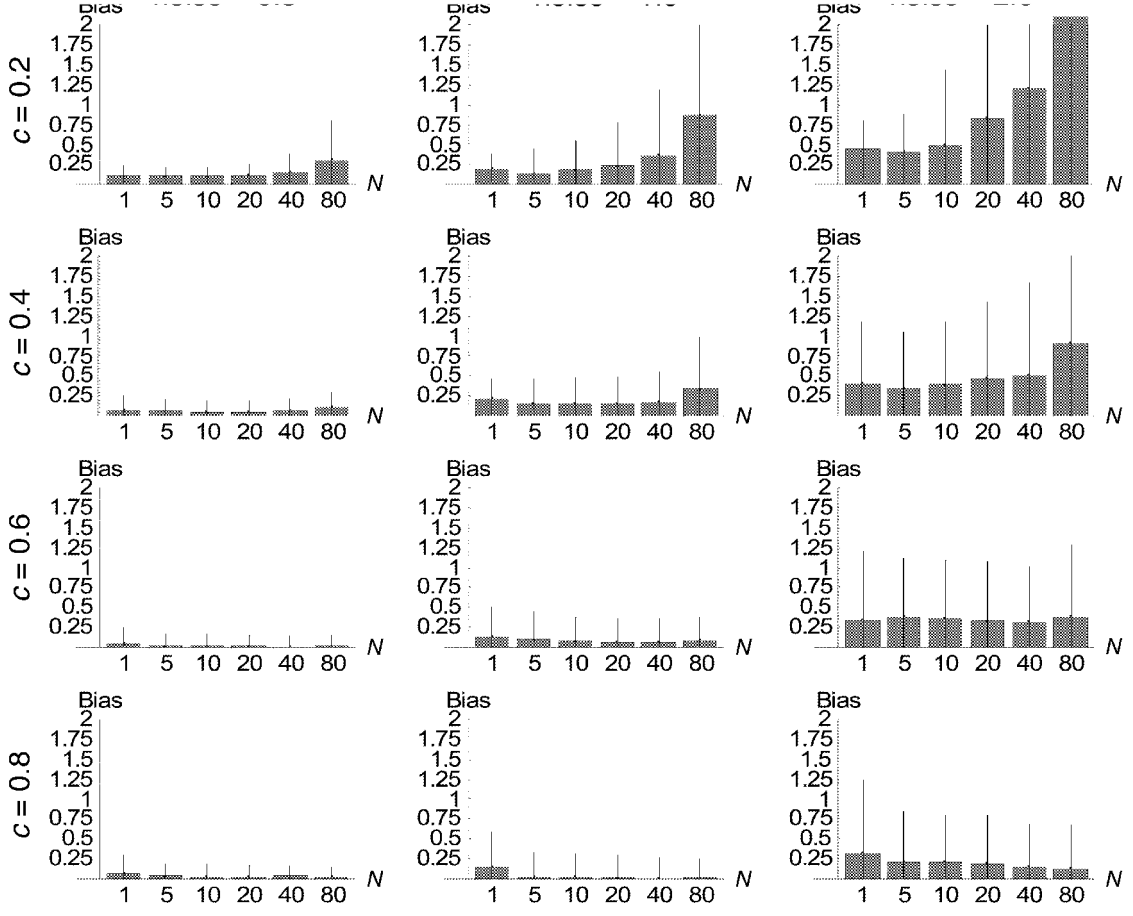

Figure A2. Bias and efficiency in percentage as a function of block size *N* for curvature *c*, increasing from top to bottom, and noise level η, increasing from left to right. A block size *N* of 1 means that no block average was used.

400 trials. This implies a kind of tradeoff, since, as a consequence, the larger the block size *N*, the fewer points remain for fitting. Everything else is as in the previous simulations. The third column of Table A1 recapitulates the fixed and varied factors.

The results are shown in Figure A2. As can be seen, for c = 0.8 (bottom row), using blocks of increasing size reduces the bias and improves the efficiency. In the best case, bias is reduced twofold, and efficiency by almost 50% (block size *N* of 80). Thus, even though there are only 5 points (400 trial data averaged by blocks of 80 trials), the parameters are recovered very accurately. However, this trend reversed for curvatures smaller than 0.5, for which averaged data return more biased and less efficient estimates. Thus, for small curvatures, the small amount of blocks (5, 10, and 20 blocks of 80, 40, and 20 trials, respectively) is very detrimental. In this case, the modeler should avoid estimating parameters on block average data.

Simulation 3: Curvature Testing

We explored the reliability of the test of linear hypothesis. In order to perform a statistical test, we first generated 100 trial data sets following a power curve. They can be seen as different subjects. As before, parameter a was fixed at 300 and b at 1,000. Parameter c varied for each half of the sets, with possible values of {0.2, 0.4, 0.6, 0.8}. When the two cs are equal, the test should not reject H_0, or else it makes a Type I error. When the two cs are unequal, the test should reject H_0, or else it makes a Type II error. The difference between the two cs is the effect size: the larger the effect size, the smaller the number of Type II errors should be. We used noise at a level η of 2.0, and the number of trials T was fixed at 400. Tests were performed with a decision level of 5%. Each test was replicated a thousand times

Figure A3 shows the results. When there were 80 blocks (N = 5), there were very few Type I errors, but the power was very low: The test almost never rejected H_0. In the opposite case (5

APPENDIX A (Continued)

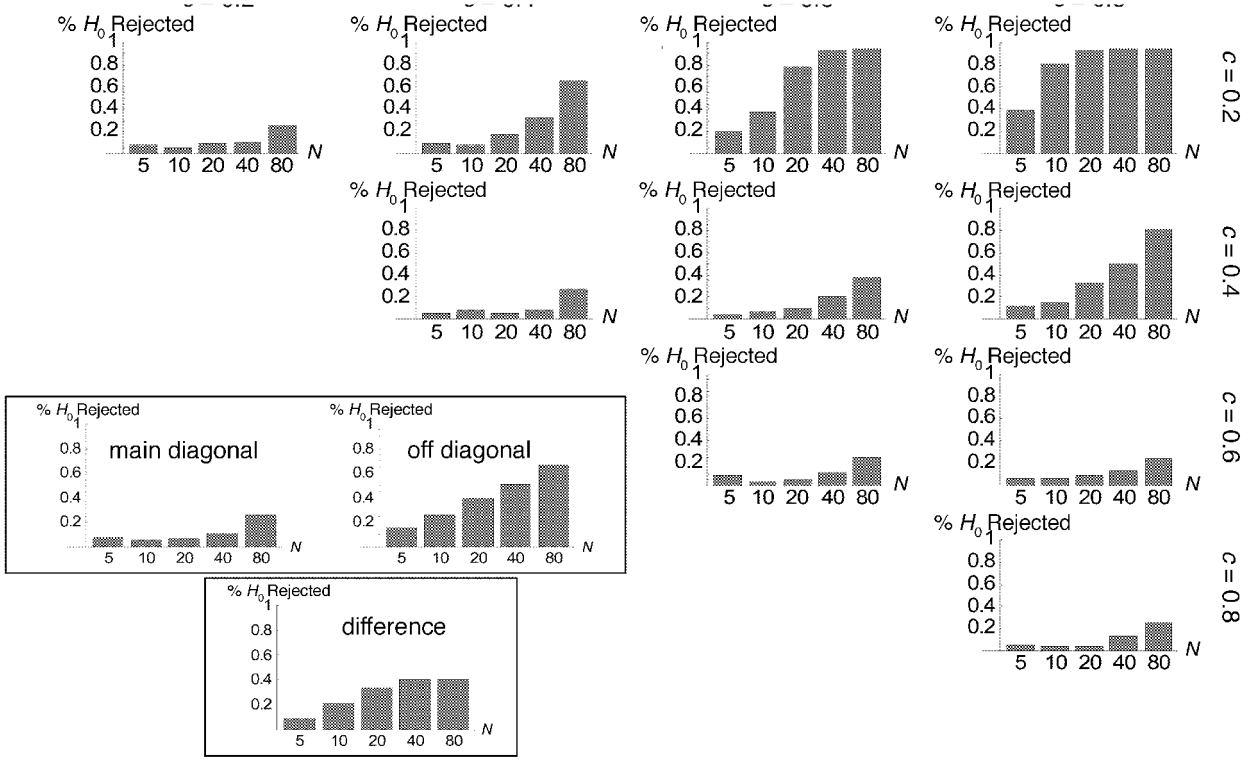

Figure A3. Proportion of times H_0 is rejected using Rao's (1959) test with 95% level of confidence as a function of block size N for curvatures of the first simulated data set, increasing from top to bottom, and curvature of the second data set, increasing from left to right. Number of trials, T, is 400, and noise, η, is 2.0. The main diagonal contains cases in which both curvatures are equal and illustrates the proportion of Type I errors. The off-diagonal plots contain cases in which curvatures are unequal and, thus, illustrate the power of the test (one minus the proportion of Type II errors). The left part of the first box shows the main diagonal across all c levels. The right part of the first box shows the power across all effect sizes. The second box shows the difference between the power and the Type I errors shown in the first box. Since this scenario weights Type I errors and power equally, the test is optimal at $N = 40$ or $N = 80$. However, if Type I errors are a concern (and are weighted more heavily), the test will be optimal at $N = 20$, the square root of the total number of trials.

blocks with $N = 80$ observations per blocks), the opposite was seen: H_0 was often rejected, resulting in a good power but a Type I error rate near 30%. Choosing the perfect compromise between block size and number of blocks (20 blocks of 20 trials) yielded the best results, with a Type I error rate near 7% and a power near 90% when a large effect size was present. Although the tests were performed with a decision level of 5%, the effective amount of Type I error was slightly larger, due to a large amount of covariation within subjects (Hoel, 1964).

In another series of simulations, we tested the efficiency of the test with 1,600 trials, and weighting Type I errors and power equally, the test was optimal at 40 trials per block, suggesting the general rule that the optimal block size for Rao's (1959) test is the square root of the number of trials.

APPENDIX B
A Mathematica Program That Performs a Test of Linear Hypothesis (Rao, 1959)

The program reads the input file "data.dat" which is composed of s columns with T observations in each. Comments are enclosed between (* and *).

```
(******* load a useful package and set working directory *******)
Needs["Statistics`MultiDescriptiveStatistics`"]
SetDirectory["C:\\WINDOWS\\Bureau\\"];

(****************** model information *****************)
Model[t_,c_] := {1,t^-c}          (* trial data power curve*)
θ [a_, b_] := {a, b}
s := 100                          (* number of columns *)

(* definition of the Sum of Square Error used for minimization *)
SSE[set_, a_, b_, c_] := Σ_{t=1}^{T} (set[[t]] - θ[a,b].Model[t,c])^2
(********************* read the data file **********************)
FileFormat = Table[Real, {s}];
data = ReadList["data.dat", FileFormat];
T = Length[data]

(****************** compute the summary values *****************)
y = Mean[Transpose[data]];
S = CovarianceMatrix[Transpose[data]];

(******* performs a fit over the average data and keep c *******)
GroupFit = FindMinimum[SSE[y, a, b, c],
    {a, 100, 300}, {b, 400, 2000}, {c, 0.2, 1.0}
  ] [[2]]
c = c/.GroupFit

(****** performs a fit for each column and average a and b ******)
IndividualFit = Table[FindMinimum[SSE[Transpose[data][[i]], a, b, c],
      {a, 100, 300}, {b, 400, 2000}, {c, 0.2, 1.0}
    ] [[2]],
  {i, 1, s}
];
θ̂ = Mean[θ[a, b] /. IndividualFit]

(******************** instantiate the model ********************)
A = Table[Model[t, c], {t, 1, T}];

(********** Perform Rao's test of linear hypothesis ***********)
r = T - Length[θ[a, b]] - 1;
n = s ;
F = (n - r)/r (y - A.θ̂) . PseudoInverse[S] . (y - A.θ̂)
```

(Manuscript received July 6, 2001;
revision accepted for publication November 10, 2002.)