Perception & Psychophysics
2001, 63 (8), 1389-1398

Slope bias of psychometric functions derived from adaptive data

CHRISTIAN KAERNBACH
Universität Leipzig, Leipzig, Germany

Several investigators have fit psychometric functions to data from adaptive procedures for threshold estimation. Although the threshold estimates are in general quite correct, one encounters a slope bias that has not been explained up to now. The present paper demonstrates slope bias for parametric and nonparametric maximum-likelihood fits and for Spearman–Kärber analysis of adaptive data. The examples include staircase and stochastic approximation procedures. The paper then presents an explanation of slope bias based on serial data dependency in adaptive procedures. Data dependency is first illustrated with simple two-trial examples and then extended to realistic adaptive procedures. Finally, the paper presents an adaptive staircase procedure designed to measure threshold and slope directly. In contrast to classical adaptive threshold-only procedures, this procedure varies both a threshold and a spread parameter in response to double trials.

A common way to analyze data acquired with adaptive psychophysical procedures for threshold estimation is to average the signal levels encountered during the run. This results in the desired threshold estimate. It is tempting to analyze the data further in order to obtain more information on the form of the psychometric function (PF). Adaptive data placement is concentrated around the threshold with a certain spread to both sides. Would the data admit an estimation of the slope (or spread) of the PF?

Adaptive data placement is intended to reduce the number of trials needed to estimate a threshold. For scarce data sets, it would be advisable to use maximum-likelihood (ML) techniques in order to obtain reliable estimates. However, several researchers have encountered the problem that there is a considerable bias in ML slope estimates from adaptive data (Leek, Hanna, & Marshall, 1992; Treutwein & Strasburger, 1999; a brief review of the empirical literature on slope estimation is found in Strasburger, 2001b). The present paper demonstrates this effect with parametric and nonparametric ML algorithms, as well as with Spearman–Kärber analysis, and then presents an explanation based on the serial dependency of adaptive data. Finally, an adaptive procedure, especially designed to estimate threshold and slope simultaneously, is presented in order to illustrate what is missing in adaptive threshold-only estimation data.

I thank Stanley Klein, Jeff Miller, Andreas Möltner, Frank Neutzler, Hans Strasburger, Bernhard Treutwein, and Dirk Vorberg for valuable discussions. Correspondence should be addressed to C. Kaernbach, Institut für Allgemeine Psychologie, Universität Leipzig, Seeburgstraße 14-20, 04 103 Leipzig, Germany (e-mail: christian@kaernbach.de).

DEMONSTRATING THE SLOPE BIAS

Slope Bias With a Parametric Maximum-Likelihood Algorithm

It is common practice with ML techniques to assume a certain PF and to vary the parameters of this function so as to maximize the likelihood of the parameter set for the run in question. This is called the parametric approach. The more that is known about the true PF, the better the ML estimates will be. Although for behavioral data the type of the true PF is not known, in simulations the ML fit can use the same type of function as do the simulated runs.

For the simulations presented in Figure 1, 10,000 runs were simulated per condition. A cumulative normal distribution was assumed to be the true PF for the simulated runs, with the working point (X_{50}) being zero and its spread $\Sigma = X_{80} - X_{20}$ being two (i.e., $X_{20} = -1$, $X_{80} = 1$; a z score unit of one is equal to $X_{84} = 1.188$). This definition of Σ corresponds to $\Sigma = X_{90} - X_{60}$ in two-alternative forced-choice tasks (for other definitions of the spread, see Strasburger, 2001a). The considered PF runs between zero and one, corresponding to the performance in a yes–no task without false alarms. The simulated adaptive runs started at the central signal level, $X_{50} = 0$. The levels were increased one step after each *no* response and increased one step after each *yes* response. Step size Δx was set to either .5 or .25. This corresponds to the maximum ($\Sigma/4$) and minimum ($\Sigma/8$) step size recommendation by Green, Richards, and Forrest (1989). The run length varied from 10 to 100 trials.

For each simulated run, an ML estimate of the slope was determined. To this end, the likelihood for this run to occur given a certain PF was calculated, and the pa-

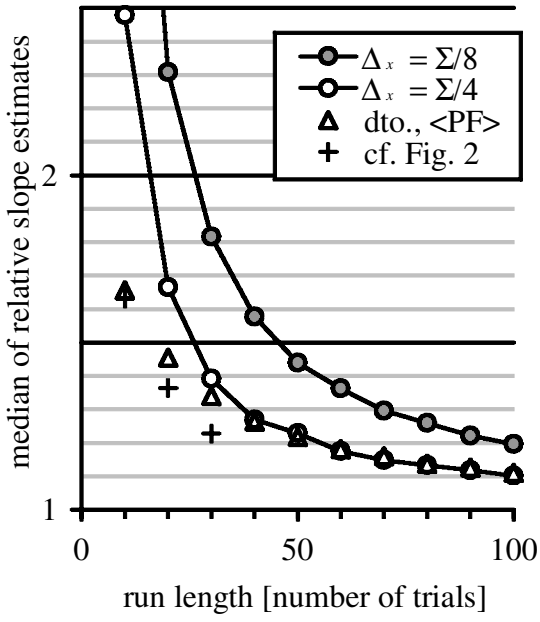

Figure 1. Demonstration of slope bias for parametric PF estimates, using cumulative normal distributions for both the simulated adaptive runs and the parametric maximum-likelihood estimates. The two curves (circles and lines) show the slope bias for two different step sizes of the adaptive staircase procedure, representing minimum and maximum recommendations for the step size. The triangles give slope bias estimates as obtained by averaging PF values instead of analyzing the slope parameters directly. The crosses refer to slope bias data from nonparametric PF estimates (see Figure 2).

rameters of the PF were varied so as to maximize this likelihood.[1] The PF was assumed to be a cumulative normal distribution, running from zero to one, with variable threshold and slope. This gave 10,000 threshold and slope estimates for each condition. Some of these slope estimates were infinite, preventing averaging of the slope estimates. Instead, Figure 1 presents the median of the slope estimates (circles), divided by the true slope. It can be seen that the median of the slope estimates is in all cases well above the true slope. This is especially so for short runs, but even for runs with 100 trials, the slope bias is around 10% (Δx = .5) or 20% (Δx = .25). The slope bias is less for the larger of the two step sizes. However, it should be noted that long runs with as many as 100 trials will in general not be performed with large step sizes throughout. When the step size is initially large, it is common practice to reduce it after a certain number of trials. Moreover, the calculation of the median underestimates the effect, since, due to the skewed distribution of the slopes, there is a large number of very high slope estimates. For instance, a quarter of all 50-trial runs with Δx = .25 yielded a slope estimate larger than twice the true slope.

Figure 1 also presents an alternative way to summarize the slope bias effect. Instead of analyzing the slope parameter directly, for each simulated run the PF values of the psychometric functions were calculated at signal intensity $x = -0.5$ and $x = +0.5$. These values were averaged across all runs, and then the slope was determined from these two values and compared with the slope of the true PF (determined in the same way from two PF values). Figure 1 shows the resulting data as triangles for step size Δx = .5 (i.e., $\Sigma/4$). For short runs, there is much less slope bias if it was determined from averaged PF values than if it was determined from the ML slope parameters directly. This is probably due to the averaging of PFs with different threshold estimates into a single average PF, thereby smoothing out the slope of this average PF.[2] For runs with 40 trials or more, there is almost no difference between these two estimates for the slope bias. Although this analysis clearly underestimates the slope bias for short runs, it has the advantage that no infinite values can occur (such as with ML slope estimates), and it can be applied to nonparametric ML fits (see next section).

Slope Bias With a Nonparametric Maximum-Likelihood Algorithm

It could be presumed that the slope bias is due to the parametric ML approach, and that there might be other estimation methods that would reliably find the correct slope. The search for such methods would, however, be strongly discouraged if one were able to demonstrate that even a nonparametric ML algorithm fails to reproduce the true form of the PF. As complicated as its name sounds, a nonparametric ML algorithm does not do much more than estimate the value of the PF at each signal intensity, given the number of correct and incorrect responses at that level. Here, the psychometric "function" need not be a specific function of known type: It is just a set of PF values at certain signal levels. The only prerequisite is that, for every single run, a PF value is determined for all signal levels in question, even if a certain level had not been tested by this run. If even this direct analysis of the "raw data" shows a slope bias, there is not much hope in finding a way to analyze these obviously biased data so as to obtain an unbiased slope estimate.

The following analysis demonstrates the slope bias effect with a nonparametric ML algorithm. This time a complete analytical analysis of all 2^N possible runs of length N (a trial can have two results, N trials can have 2^N results) is performed. In contrast to simulations, the results of such a combinatorial approach can be considered exact. Due to the exponential increase in computation time, the maximum run length tested is $N = 30$. A PF was chosen that was identical to the PF of the previous simulation; the step size Δx was set to $\Sigma/4 = 0.5$ (i.e., the value that had shown better slope estimates with the parametric ML algorithm). Let us assume that we perform adaptive runs of N trials, with the signal level x starting at $X_{50} = 0$ and increasing one step after each *no* response and decreasing one step after each *yes* response. Some of the 2^N runs are less likely, and others are more likely. The likelihood for each of these 2^N runs was calculated. Furthermore, the PF that would have resulted from each run was calculated. The average of the

ML estimates for each signal level, weighted with the likelihood of the respective run, can then be compared with the true PF.

The value of the PF at each signal level was first assumed to correspond to the proportion of correct responses from the total number of tests at this signal level. This PF is not necessarily monotonically increasing. The only deviation from using the raw data was to make the data of each run monotonic. This is called *isotone regression*, and the algorithm is called *pool adjacent violators* (PAV; see, e.g., Barlow, Bartholomew, Bremner, & Brunk, 1972). This algorithm pools the data sets from signal levels that violate the monotonicity (for a more detailed explanation of the algorithm, see also Miller & Ulrich, 2001). There are two reasons for applying the PAV algorithm: First, such deviations from monotonicity would in general not be thought of as true features of the PF, and it would be appropriate to apply a PAV algorithm before presenting one's data, and especially so if one is interested in the crossing of the PF with a predefined threshold probability. Otherwise one might end up with more than one crossing of the raw PF and threshold probability. Second, and more importantly, the monotonicity constraint improves the estimates at the borders of the tested signal range where only few tests occur and admits to estimate the values of the PF at those signal levels that have not been tested (i.e., above or below the tested range). For the present approach this extrapolation is necessary since the averaging of the PF values can only be performed if all runs yield PF estimates for all signal levels in question. It should be noted that the monotonicity constraint is not specific to this approach: Parametric ML algorithms fit monotonic functions to the data, thus imposing monotonicity (and further constraints).

The monotonicity constraint leaves a certain range of possible values for the PF values at levels not tested; if, for instance, the leftmost value of the monotonized PF is 0.1, the values to the left of it can be anywhere in the range of zero to 0.1. This *continuation uncertainty*, however, occurs outside the center of the PF (where the slope is measured) and is not important for adaptive procedures. Figure 2A shows the two extreme interpretations ("conservative," i.e., taking the most central, close to 0.5, interpretation and "asymptotic," i.e., assuming the asymptotic values for all values that have not been tested) for $N = 10$. For higher values of N, the continuation uncertainty gets even smaller.

Figure 2B shows the average of the conservative estimates of the monotonized raw data for $N = 10$, 20, and 30, as compared with the true PF. Obviously, the general form of the obtained PF cannot be relied on. The estimated PF is steeper than the original one. For example, with $N = 10$, the true PF at the intensity level $x = -1$ is 0.2, whereas the estimated PF is 0.08 (conservative estimate).

Nonparametric estimates do not provide direct values for the slope parameter. The slope bias can be estimated, however, from the linear interpolation of the signal levels neighboring the central level as compared with the same interpolation of the true PF. These slope bias estimates are shown as crosses in Figure 1. They should be compared with the parametric slope bias estimates obtained by averaging the PF values (triangles in Figure 1); they are smaller than those estimates for runs of more than 10 trials.

Slope Bias With Stochastic Approximation

It could be suggested that the slope bias is due to the rigid staircase type of procedure employed hitherto. In computer simulations, more flexible schemes of data

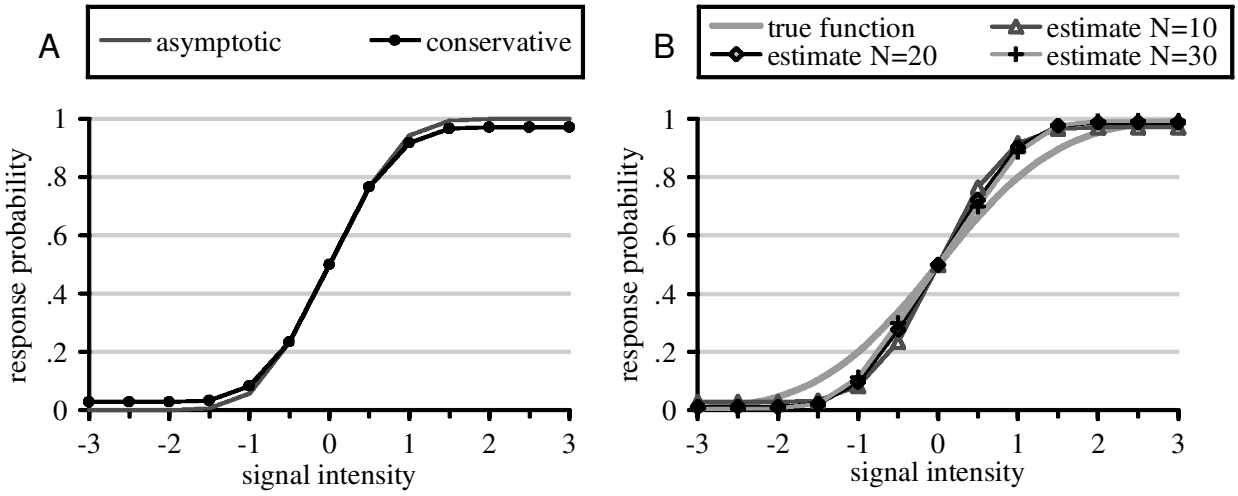

Figure 2. Nonparametric maximum-likelihood estimates of the psychometric function. (A) Comparison of the asymptotic and the conservative estimates. The difference between these two curves stems from the undeterminedness of estimates at signal levels that were not tested, limited only by the monotonicity constraint. (B) Nonparametric conservative maximum-likelihood estimates for different run lengths, as compared with the true psychometric function. The slope bias derived from linear interpolation of these data (−0.5 to 0.5) and comparison with the same linear interpolation of the true psychometric function is shown in Figure 1 as crosses.

placement, such as the stochastic approximation algorithm (Robbins & Monro, 1951), produce less statistical fluctuation of the threshold estimates with the same number of trials. In behavioral experiments, most researchers, even when using staircase procedures, alter the step size once or twice during a run, thereby emulating stochastic approximation. Could it be that this procedure would also yield more reliable slope estimates?

For the following demonstration, the same analysis was done as for Figure 2B, the only difference being that the data placement followed a rule of stochastic approximation. The starting level was again at $X_{50} = 0$, and after each *yes* response the signal level was increased by one step size Δx_t, and after each *no* response it was decreased by Δx_t. In contrast to the previous analysis, the step size Δx_t was not fixed but was a function of the trial number t. In stochastic approximation, the step width is usually reciprocally related to the trial number: $\Delta x_t = D/t$, with a constant D. To keep the difference between the first few step widths within a reasonable range usually the first trials are skipped, or in other words, a constant value is usually added to the denominator: $\Delta x_t = D/(t+t_0)$. In the present analysis, t_0 was set to 3, and the constant D was chosen in order to give average step sizes across the entire run of N trials that were comparable to the value that produced the smaller slope bias in Figure 2B ($\Delta x = 0.5$). The following combinations were used: $N = 10$: $D = 3.7$; $N = 20$: $D = 5.2$; $N = 30$: $D = 6.6$. As with Figure 2B, a PAV procedure was used to calculate the ML estimate that would result from each possible run. The average of these estimates was weighted by the likelihood for the respective run (i.e., the expectation values of the estimates are shown in Figure 3).

The curves for different run lengths coincide nearly perfectly. For $N = 10$ there are some deviations visible that are due to the coarse graining of the still large steps after 10 trials. For larger run lengths, the estimated PFs get quite smooth and can be perfectly matched by a cumulative normal distribution. This is remarkable since the ML approach used in this analysis did not specify the type of function that was used as the true function. The estimated PFs are, however, quite precisely doubly as steep as the original PF.

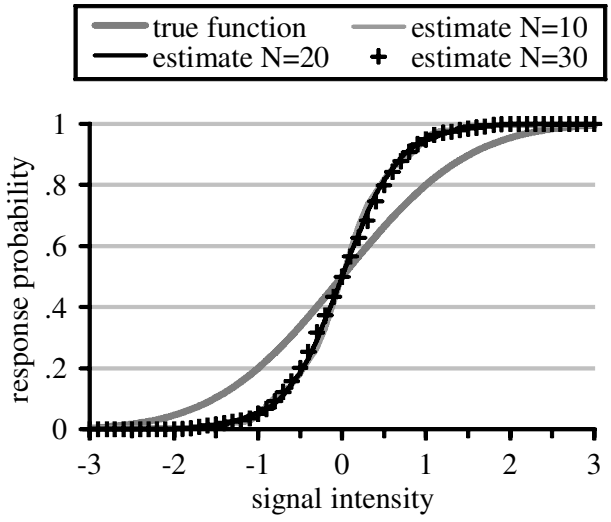

Figure 3. The same analysis as for Figure 2B, but using stochastic approximation instead of a staircase procedure. This time, the slope bias seems not to decrease with increasing run length.

By comparing Figure 3 with Figure 2B, it can be noted that for staircase procedures (with constant step width), there is, with increasing N, an asymptotic approach of the estimated functions to the original function, whereas for the stochastic approximation, the slope bias is independent of N. The ever decreasing step size of the stochastic approximation method seems to favor the slope bias, whereas the constant step width of staircase procedures has the effect that, for increasing run lengths, levels far from the working point are tested again and again (see section on data dependency).

Slope Bias With the Spearman–Kärber Method

ML estimation techniques are often considered the optimum analysis for probabilistic data. They require independent data gathering (i.e., the number of trials per intensity level should be fixed and should not depend on the result of previous trials at this or any other level). If this precondition is not met, the results of the ML analysis are prone to bias. It could thus be conceived that non-ML techniques are superior to ML with regard to the slope analysis of adaptive data. In the this issue of *Perception & Psychophysics*, Miller and Ulrich (2001) evaluate the Spearman–Kärber method, and find it superior in cases when the ML assumptions are not met. In addition to a reduced systematic error, as compared with probit analysis, the method gives information on higher moments of the distribution such as skewness and kurtosis. However, the authors evaluated this method with simulations of constant-stimulus data. Therefore, it was not evident whether there would be slope bias when adaptive data were analyzed with this non-ML technique.

We tested whether a Spearman–Kärber analysis of adaptive data would show the slope bias that was found for ML techniques.[3] A set of 10,000 simulated adaptive simple up–down runs of 30 trials each (same psychometric function as for Figure 1; step size $\Sigma/4$) was analyzed with the Spearman–Kärber method and with ML probit analysis. The median standard deviation for the ML probit analysis was 0.86. For the Spearman–Kärber method, the median standard deviation was 0.71. Given that the standard deviation of the true psychometric function was 1.188, this translates to slope ratios of 1.188/0.86 = 1.38 for the ML analysis (compare also Figure 1) and of 1.67 for the Spearman–Kärber method. The slope bias found with the Spearman–Kärber method is nearly twice as large as that of the ML probit analysis. A possible explanation is that the Spearman–Kärber method weighs all levels equally, whereas the ML techniques weigh the levels according to the number of trials and hence, quite appropriately, give less weight to the border levels that contribute more to the slope bias.

EXPLAINING THE SLOPE BIAS

Whereas the previous sections have demonstrated slope bias under various conditions, the following sections attempt to explain its origin. It will be shown that this bias is due to the fact that adaptive data have not been collected independently. Independent data gathering would require that the experimenter determine in advance the levels to be tested and the number of tests to be performed at these levels. Adaptive data are different: The output of previous trials determines the if and where of subsequent trials. It is incorrect to treat such data the same way as constant-stimuli or other independent data, by looking at the outcome of each individual trial and trying to learn the most of it about the PF. Their information is in the stimulus placement, and once this information has been evaluated (by averaging signal levels), they are—to put it simply—"wrung out" and should not be analyzed further. Not even ML techniques can extract any further information.

The first section below demonstrates that the often suspected distribution of the tests is not at the origin of slope bias. The next two sections illustrate the effects of data dependency with simple two-trial examples. With adaptive data, the dependency of the data is multiple and complex, and so are the effects of this dependency. The outcome of one trial will bias the retest probability at this same signal level, as well as the test probabilities of other signal levels (with the results of these tests rebiasing the test probabilities of the original level, and so forth). Although the effect of the entire network of dependencies is demonstrated in Figure 2B (please remember that this is an exact calculation), for the purpose of illustration it is helpful to focus on *elementary dependencies* as demonstrated in these sections with two-trial examples. The last section extends the argument to adaptive sequences of realistic length.

Effect of the Distribution of the Tested Signal Levels

The slope bias is not due to the distribution of the signal levels that are tested. It is often suspected that the focusing of the testing to the center of the PF (and, in consequence, the sparse data collection at positions that seem more important to slope estimation) is at the origin of the slope bias. This is not so.

Andreas Möltner (personal communication, July 26, 1995; cf. Pflug, 1990) suggested that if one were interested in determining the slope of the PF while using adaptive procedures, one should use a kind of duplex procedure, combining the data placement of adaptive procedures with the independent testing of a constant-stimulus approach. After each trial of the adaptive procedure, one should do a second trial at the same level (i.e., before changing the level as required by the procedure) and only use the data of this second trial. The adaptive procedure would determine where the tests are performed, whereas the response data to enter the ML algorithm would have been obtained independently of the correctness of the previous response. This is a useful demonstration of where the slope bias effect does not come from.

Figure 4 shows nonparametric ML estimates of the PF using the same staircase procedure for stimulus placement that was used in the previous section. Again, a complete analysis of all 2^N possible runs was performed. This time, both conservative and asymptotic estimates are shown (cf. Figure 2A). At high and low levels, they differ much more for the adaptively placed constant stimuli data than they do for the adaptive data (Figure 2A). The uncertainty as to the PF values outside the tested range reflects the intrinsic effect of the data-focusing achieved by the adaptive data placement. This focusing does not, however, imply slope bias: The two types of estimates envelop the true PF from both sides, and for higher N, they do so more closely, without any systematic bias of the slope. Already for $N = 10$, the range of possible estimates at the levels neighboring X_{50} ($x = -0.5$: [0.30,0.36]) is close and centered around the true value (0.33), whereas the estimate for adaptive data (0.23) is well below this range.

It is interesting to consider why, for adaptive data, the conservative and the asymptotic estimates differ much less than for those adaptively placed constant stimuli data. In the latter case, the difference reflects the range of outcomes at the bordering levels. With adaptive data placement, the tests at these levels are used for placing further trials. If the termination criterion is given in reversals, the leftmost test will necessarily have had a negative result. Otherwise it would not have been the leftmost test. If the termination criterion is given in absolute trial number and not in reversal number, it might occur that the run ended at the leftmost level tested and the test

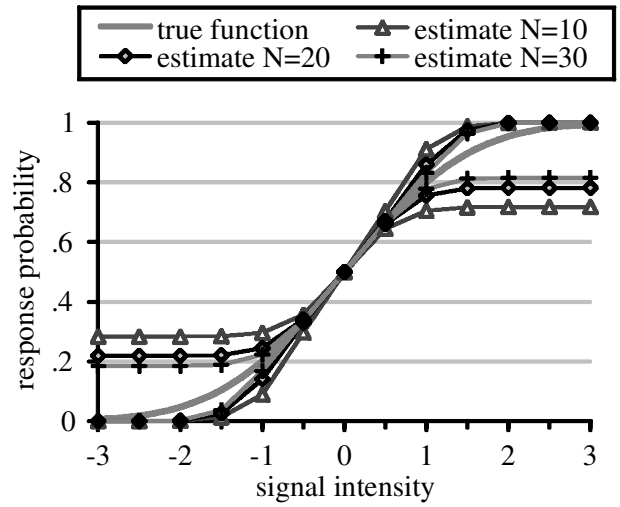

Figure 4. The same analysis as for Figure 2B, but with adaptively placed constant-stimuli data, for run length $N = 10, 20$, and 30 trials (see text for explanation). The steeper curve for a given run length shows the asymptotic estimate, the shallower curve shows the conservative estimate (see Figure 2A). There is no slope bias in these data: The true PF is enclosed between asymptotic and conservative estimate.

there was positive. These rare cases are at the origin of the small difference between the asymptotic and the conservative estimates observed in Figure 1A.

Retest Probability for a Single Level

The problem that one encounters while trying to analyze dependent data is best illustrated with extremely short sequences of one or two trials. Although in adaptive sequences successive trials are usually at different levels, let us, for the moment, consider that there is only one level under question. Let us first consider ML estimates of independent data. Imagine a sequence of n trials, each with two possible outcomes (positive: $\oplus$, negative: $\ominus$). If m trials have a positive result, and if the a priori distribution of possible probabilities is uniform, the ML estimate e of $p_\oplus$ is equal to m/n. Imagine one performs a single trial. In case of a positive result, the ML estimate for $p_\oplus$ is $e_\oplus = 1$, and after a negative result, it is $e_\ominus = 0$. Due to the limited information that can be gained from a single trial, these two-valued estimates are only rough estimates of the probabilistic quantity $p_\oplus$. They are, however, not biased. This can be seen by considering the expectation value of the ML estimate e:

$$<e> = p_\oplus \cdot e_\oplus + p_\ominus \cdot e_\ominus = p_\oplus \cdot 1 + (1-p_\oplus) \cdot 0 = p_\oplus. \quad (1)$$

The same is true if one performs two trials. In this case there are four possible results, namely $\oplus\oplus$, $\oplus\ominus$, $\ominus\oplus$, and $\ominus\ominus$. The ML estimates for $p_\oplus$ as a function of these results are $e_{\oplus\oplus} = 1$, $e_{\oplus\ominus} = e_{\ominus\oplus} = 0.5$, and $e_{\ominus\ominus} = 0$. The expectation value for e is again correct:

$$\begin{aligned} <e> &= p_{\oplus\oplus} \cdot e_{\oplus\oplus} + p_{\oplus\ominus} \cdot e_{\oplus\ominus} \\ &\quad + p_{\ominus\oplus} \cdot e_{\ominus\oplus} + p_{\ominus\ominus} \cdot e_{\ominus\ominus} \\ &= p_\oplus^2 \cdot 1 + 2 \cdot p_\oplus \cdot (1-p_\oplus) \cdot 0.5 \\ &\quad + (1-p_\oplus)^2 \cdot 0 = p_\oplus. \end{aligned} \quad (2)$$

The problem starts if one decides to perform a second trial only if the first trial had a certain result. Imagine, for example, a teacher who would perform a second test only if the first test had a positive result. The students would claim that if they encounter the risk to fail after a positive test, they should as well have a chance to improve after a negative test. This *biased test* has three possible outcomes: $\oplus\oplus$, $\oplus\ominus$, and $\ominus$. The ML estimates for these results are $e_{\oplus\oplus} = 1$, $e_{\oplus\ominus} = 0.5$, and $e_\ominus = 0$. The expectation value for e is biased toward too small values:

$$\begin{aligned} <e> &= p_{\oplus\oplus} \cdot e_{\oplus\oplus} + p_{\oplus\ominus} \cdot e_{\oplus\ominus} + p_\ominus \cdot e_\ominus \\ &= p_\oplus^2 \cdot 1 + p_\oplus \cdot (1-p_\oplus) \cdot 0.5 + (1-p_\oplus) \cdot 0 \\ &= p_\oplus - p_\oplus(1-p_\oplus) \cdot 0.5 < p_\oplus, \text{ for } 0 < p_\oplus < 1. \end{aligned} \quad (3)$$

With the opposite rule (perform second trial only if first trial was negative), the effect will be inverse: The expectation value will be biased toward high values.

There is no way out of this dilemma. The only possible choice of estimates e that would give a correct expectation value for the estimate would be to set $e_{\oplus\oplus} = e_{\oplus\ominus} = 1$. With this choice of non-ML estimates, one would completely disregard the result of the second trial. One would have to assume a probability $p_\oplus$ of 1 even if the result was $\oplus\ominus$, which is obviously wrong. In the case $\oplus\ominus$, we have more information. We know that we should not assume $p_\oplus$ to be 1, but we know that by using this information, we are prone to obtain on the average a biased result.

Let us assume that the rule is less strict than in the above example. Let the probability of a retest after a positive result in the first trial be $r_\oplus$, and that for a retest after a negative result be $r_\ominus$. The formula for $<e>$ will then have to deal with six possible outcomes ($\oplus$, $\ominus$, $\oplus\oplus$, $\oplus\ominus$, $\ominus\oplus$, $\ominus\ominus$). From a calculation similar to Equation 3, it follows that

$$<e> = p_\oplus - (r_\oplus - r_\ominus) \cdot p_\oplus(1-p_\oplus) \cdot 0.5. \quad (4)$$

Only if the retest probability does not depend on the outcome of the first trial (i.e., $r_\oplus = r_\ominus$), will the expectation value of the estimate be correct.

The Test-at-All Probability for Two Trials Involving Three Levels

The last section dealt with two trials that can be performed at the same level. It is more typical for an adaptive procedure that a second test would take place at a different level. Let us consider a very simple sequence of two trials of a simple up–down run. After a positive trial at level L, the next trial is placed at level L−1, after a negative trial it is placed at level L+1. The two-trial sequence will have four possible outcomes. Two of these outcomes imply tests at L and L−1, and the other two runs will test levels L and L+1. Let us consider the estimate at level L−1. If tested, the outcome will depend on the value of the PF at this level. If not tested, this is due to a negative trial at level L. The ML estimate for level L is zero, and assuming monotonicity, the estimate for L−1 also has to be zero.

The situation is similar to that of the biased two-trial test at the same level that led to Equation 3. There, after a negative trial at a certain level, no further trial was performed. Instead, the outcome of a second test at the same level was anticipated to be negative as well. With the present biased two-trial test, a *first* test at level L−1 is not considered after a negative test at level L. Again, it is anticipated to be negative. In contrast to Equation 3 the calculation of the average estimate of the PF at level L−1 involves two levels and their respective probabilities:

$$<e(\mathrm{L}-1)> = p_\oplus(\mathrm{L}-1) \cdot p_\oplus(\mathrm{L}) < p_\oplus(\mathrm{L}-1). \quad (5)$$

Note that this is not a specialty of isotone regression: Parametric ML techniques impose even stronger constraints on the form of the PF, including monotonicity.

Data Dependency in Adaptive Psychophysical Sequences With More Than Two Trials

The slope bias observed when analyzing the data of adaptive procedures of a realistic length can be attributed to both of the sources mentioned above:

The retest probability for a signal level below the equilibrium point is higher after a positive result than after a

negative result in the previous result at that level. The reason for this is easy to see: The average drift at intensity levels below the equilibrium point is positive (i.e., given that the run is now at this intensity level) and it will probably soon be at higher intensity levels. A positive result entails a movement against the current, and so will soon lead to a retest. A negative result, on the other hand, leads to intensity levels around the equilibrium point where the run undergoes stochastic random walk processes. In line with Equation 4, this will lead to an underestimation of the PF value at this level. The same reasoning will yield overestimation of the PF values above the equilibrium point.

The test-at-all probability for a signal level below the starting point is higher after a positive result at its right neighbor than after a negative result. This is due to the nature of adaptive rules. Given that the estimate will be zero if this level is not tested at all, this will lead to an underestimation of the PF values at this level. The test-at-all effect may be noneffective on one half of the PF, if the starting point is chosen well apart from the true threshold. In most psychophysical experiments, the starting point is well above threshold, and, in this case, the levels above the equilibrium points are tested during the initial phase of the run. The situation is, however, probably different if the step size is altered during the run; in this situation, not all levels above the equilibrium point will be tested during the initial phase.

The quantification of the respective effects of these two sources of slope bias is difficult. In longer adaptive runs, Equation 4 does not apply. With 10 trials starting at level 0 (step size 0.5), the level at $x = -1$ can be tested four times. The probability for that level to be tested a third (fourth) time depends not only on the outcome of the first trial, but also on that of the second (and third) trial. As compared with the 6 possible outcomes that enter Equation 4, this time there are 30 ($2^1 + 2^2 + 2^3 + 2^4$) possible outcomes, and their respective probabilities and resulting estimates need to be considered. The effect of the test-at-all probability also involves a much more complicated formula than Equation 5. The total effect of data dependency can be calculated (cf. Figure 2), but it appears to be difficult to disentangle the relative contributions of retest probability and test-at-all probability.[4]

ADAPTIVE THRESHOLD AND SLOPE ESTIMATION

Recently there have been suggestions for Bayesian adaptive procedures that estimate the slope directly (King-Smith & Rose, 1997; Kontsevich & Tyler, 1999) instead of deriving it from data from adaptive threshold estimation procedures. Adaptive slope estimation is not in contradiction with the tenor of the present paper. In these Bayesian adaptive slope estimation procedures, a two-dimensional array of threshold and slope parameters is maintained, and the value of the next trial is selected in a Bayesian manner in order to obtain maximum information on the parameters. In contrast to that, a threshold estimation procedure maintains only a threshold estimator, and slope estimation from the data of such procedures is prone to bias.

In this section, I present another adaptive procedure especially designed for the simultaneous estimation of threshold and slope. It is simpler than the Bayesian approaches, modifying threshold and slope estimators according to adaptive rules comparable to those for the threshold estimator in classical staircase procedures. Its purpose is mainly to serve as an illustration of how an adaptive procedure for slope estimation should operate. From this it should then be clear why a classical adaptive procedure just does not collect the type of data that is needed for slope estimation.

Adaptive threshold and slope estimation (ATASE) is based on double trials at two different signal levels, a high (H) and a low (L) one. The distance $\Sigma = H - L$ is an estimator for the spread (i.e., the reciprocal of the slope) of the PF, whereas the mean $T = (H+L)/2$ is an estimator of the threshold. There are four possible outcomes in each trial. If both tests yield identical results, the threshold estimator should be changed, by moving it either up (in case of two negative results) or down (positive results). If the two tests of a trial yield different results, and if these are in accordance with the position of the tested levels (i.e., L$\ominus$ and H$\oplus$), it is possible that the spread is smaller than the spread estimator, and the spread estimator should be lowered. If the two tests yield results that seemingly do not accord with their positions (L$\oplus$H$\ominus$), this is an indication that the spread estimator is too small and that the probabilities for $\oplus$ are comparable at H and at L, occasionally resulting in L$\oplus$H$\ominus$. In this case, it is advisable to enlarge the spread estimator.

Overall, eight level adaptations have to be specified (ΔL and ΔH, for all four possible outcomes). In order to reduce the number of the degrees of freedom, some reasonable constraints can be imposed. The two major constraints are that the procedure converge to the desired estimates for the threshold and the slope. To this end, the average drift of the threshold and of the slope estimator can be calculated, given the probabilities p(H$\oplus$) and p(L$\oplus$) for positive outcomes at the two signal levels, and given the eight level adaptations. The two constraints consist then in setting this drift equal to zero for the desired probabilities p(H$\oplus$) and p(L$\oplus$). Four further constraints can be imposed by demanding that either the slope or the spread parameter is altered, but not both at once. This would, for example, imply that $\Delta L_{L\oplus H\ominus} + \Delta H_{L\oplus H\ominus} = 0$, in order to have no change of the threshold estimator in case of the L$\oplus$H$\ominus$ event (in which a change of the spread parameter is mandatory). This leaves two degrees of freedom, corresponding to the two step sizes for slope and spread convergence.

Table 1 gives a possible reaction scheme, leading to X_{50} threshold estimates and $X_{67} - X_{33}$ spread estimates. Two positive or negative results are treated as they would be in a simple up–down procedure, going one step up of

Table 1
Adaptive Threshold and Slope Estimation (ATASE) for Convergence to $T = X_{50}$ and $\Sigma = X_{67} - X_{33}$

Result	Probability at Target Points	ΔL	ΔH	ΔT	$\Delta\Sigma$
L⊕H⊕	2/9	$-\alpha$	$-\alpha$	$-\alpha$	0
L⊕H⊖	1/9	-4β	$+4\beta$	0	$+8\beta$
L⊖H⊕	4/9	$+\beta$	$-\beta$	0	-2β
L⊖H⊖	2/9	$+\alpha$	$+\alpha$	$+\alpha$	0
Net effect				0	0

Note—ΔL and ΔH are the adjustments of the low and the high signal level, respectively, in response to a certain result of the double trial (first column). ΔT and $\Delta\Sigma$ follow from $\Sigma = H - L$ and $T = (H + T)/2$. The net effect (last row) is calculated by multiplying the respective column with the probability column (second column) and summing up.

size α in the negative case, and one step down of the same size in the positive case. The spread estimator Σ is not affected in these cases. Mixed results will not alter T, since the movements of L and H are in opposite directions and of the same size. Here, Σ is either enlarged (if unexpectedly the test at L was the positive test) or reduced. Assuming a symmetric PF, and given that T is approaching X_{50}, the probabilities for H⊕ and L⊕ will add to 1. Enlargement of Σ will then occur with $(1-p)^2$, and reduction with p^2, with p being the probability for L⊕. The ratio between enlargement and reduction should then be equal to $(1-p)^2/p^2$ (which is, e.g., 4 in case of $p = 1/3$). The last line of Table 1 shows the drift of the two estimators at the target points, obtained by multiplying the probability column and the respective adjustment column and adding across all four possible results. They are both 0, which is compatible with the two major constraints on the eight parameters—namely, that the procedure converges to the desired target points. Note the four zero entries in the T and Σ columns, corresponding to the four minor constraints of independence of T and Σ variation.

The remaining two parameters α and β can be chosen appropriately so that the convergence process is fast. A possible advantage of the ATASE procedure is that one disposes of an on-line estimator of the spread of the PF. It would appear appropriate to base both α and β on the actual value of this spread estimator (i.e., on Σ). Instead of prescribing adaptations of L and H, one can prescribe manipulations of Σ and T, which is equivalent. Since the spread estimator Σ should not become negative, the widening and shrinking of the spread could be done by multiplication instead of by addition. This would be equivalent to varying $\log(\Sigma)$ by additive increments.

There are many issues to consider before employing this procedure, such as good starting values, step sizes (α and β), change of step sizes during the run, termination criteria, and data analysis issues such as the proper amount of discard. The present paper will not elaborate on these issues. The purpose of introducing ATASE was to demonstrate what kind of data an adaptive procedure should collect to be able to estimate the slope of the PF. The important difference to classical adaptive procedures is that, here, a spread parameter is varied adaptively, and its value can be averaged across the trials of the adaptive run. For practical purposes, one might choose a different approach. Table 2 presents eight level adaptations that fulfill the two major criteria (convergence to the target points, see last row of Table 2) but violate two of the independence constraints: For identical positive or negative results, the spread changes. This set of level adaptations fulfills another type of independence: The adaptations of L do not depend on the results at level H, and vice versa. In other words, the procedure described in Table 2 is nothing more than interleaving two classical adaptive staircase procedures (here: weighted up–down; cf. Kaernbach, 1991) with two different convergence levels (here: 1/3 and 2/3). After H⊕, level H is decreased one step, and after H⊖, it is increased two steps; this leads to a convergence point with $p(H⊕) = 0.67$. For L, the opposite rule is applied, leading to a convergence at X_{33}. The advantage of this approach is that much is known about classical adaptive procedures for threshold estimation. Levitt (1971) had already proposed to interleave two runs, aiming at different target points of the PF in order to obtain spread information. On the other hand, during the procedure, no use is made of the spread estimator. Whether or not the independence constraints realized in Table 1 or the availability of an on-line spread estimator for step size adjustment pay off in any advantage such as smaller and/or independent errors of the threshold and slope estimates remains to be seen.

It is interesting to compare ATASE and the Bayesian approaches by King-Smith and Rose (1997) and by Kontsevich and Tyler (1999) with classical adaptive threshold estimation procedures. Both the rule-based approach of ATASE and the Bayesian approaches differ from classical threshold estimation in the fact that a slope estimator is evaluated during the run. It seems to be the lack of this feature that prevents bias-free slope estimation from classical adaptive data. Whereas ATASE performs blocks of two trials at a certain distance, the Bayesian approaches perform single trials with a bimodal distribution of test levels. It is yet unclear how important this difference is. Slope estimation is not yet as well studied as

Table 2
Interleaving Two Adaptive Weighted Up–Down Runs for Target Points X_{67} and X_{33}

Result	Probability at Target Points	ΔL	ΔH	ΔT	$\Delta\Sigma$
L⊕H⊕	2/9	-2α	$-\alpha$	$-3\alpha/2$	$+\alpha$
L⊕H⊖	1/9	-2α	$+2\alpha$	0	$+4\alpha$
L⊖H⊕	4/9	$+\alpha$	$-\alpha$	0	-2α
L⊖H⊖	2/9	$+\alpha$	$+2\alpha$	$+3\alpha/2$	$+\alpha$
Net effect				0	0

threshold estimation. It would be highly interesting to compare the different approaches (including the approach to measure two different points of the PF) with regard to their efficiency and reliability.

DISCUSSION AND CONCLUSIONS

The present paper has demonstrated slope bias effects when the slope was estimated from data of adaptive threshold estimation procedures. Slope bias was found for staircase procedures with constant step size, as well as for a stochastic approximation algorithm with decreasing step size. Parametric ML fits showed slope bias, as did nonparametric ML fits and Spearman–Kärber analysis. Slope bias with adaptive threshold estimation data seems to be a general phenomenon and not to be restricted to certain adaptive methods or analysis methods.

Two sources of this slope bias have been identified and illustrated with simple two-trial examples: The *retest probability* at a certain level depends on the results of previous trials at this level, and the probability to test a certain level at all (the *test-at-all probability*) depends on the results of trials at the neighboring levels. These two mechanisms should not be considered to be different in nature, but to be representing two aspects of the same source for slope bias (i.e., of serial dependency of data).

It is often suspected that the slope bias is due to the uneven distribution of the test levels: Adaptive procedures focus tests in the threshold region and thus give only sparse information on those regions of the PF that might be of interest to slope estimation. The present paper demonstrates that this is not so. Constant-stimulus data that show the same uneven distribution of tested levels do not give rise to slope bias (see Figure 4 and corresponding text). Only if the data that served for stimulus placement are used at the same time for slope estimation will slope bias occur. This is a clear indication that slope bias is due to data dependency.

It is difficult to quantify slope bias exactly. In simulations, one can determine the distribution of the slope estimates of parametric ML fits and compare it with the true slope of the underlying PF. It is, however, not clear which comparison represents the best measure for slope bias. The distribution of slope parameters cannot simply be averaged, since some ML slope estimates are infinite. The median of this distribution will underestimate the effect due to the skewness of the distribution. Averaging PF values instead of PF parameters will (at least for short-run lengths) underestimate the effect even more, because PFs with different threshold estimates will smooth out the slope of the average PF. On the other hand, this approach is useful since it can be applied to nonparametric ML fits. Moreover, it is more accessible to attempts to explain slope bias because the probabilities in question are subject to equations like Equations 1–5, whereas the connection to parametric ML slope estimates is more complicated. Both measures (median of slope parameters, and slope determined from averaged PF values), however, indicate clearly that there is a slope bias when determining the slope from adaptive threshold data.

For longer runs, there is less slope bias for nonparametric ML fits than for parametric ML fits (see Figure 1). This difference is not well understood. If confirmed, this finding would indicate that it would be better to estimate the slope from the monotonized raw data than from the parametric fits. It should be noted, however, that this would at best reduce the problem, not solve it.

Slope bias is larger for smaller step sizes (or, as Leek et al., 1992, put it, for a constant step size, slope bias increases with lower true slope of the PF). With constant step size, the amount of slope bias decreases with increasing run length. However, long adaptive runs are in general performed with small step sizes, at least during their later parts. In the same line, data from stochastic approximation show a constant slope bias that does not decrease with increasing run length (see Figure 3). The run lengths needed to get slope estimates with less than 10% error are rather high. Leek et al. (1992) suggest runs of at least 200 trials. For shorter runs, they suggest correction factors. Given that it is not the ML technique that is to blame but the data, it does not seem possible to correct for the bias by simply multiplying with a factor. Consider the discussion following Equation 3: The only possible correction that would avoid bias was to disregard the result of those trials that were performed or not, depending on the outcome of earlier trials. In other words, the bias can be overcome only by disregarding that part of the data that is dependent on earlier data. The correction factors may work for some sets of PF types and step sizes and may not work for other combinations.

If the slope of the PF is of interest, one can measure two different points of the PF by performing two runs with different target levels. The slope could then be determined by determining the two different target points with classical methods and calculating the slope from these values. The present paper suggests a different approach to determining threshold and slope simultaneously, by performing double trials at two different levels. In contrast to adaptive threshold-only estimation, two estimates are updated from double trial to double trial, one for the threshold and another one for the slope (or spread) of the PF. Further studies could reveal whether any advantage can be derived from on-line spread estimation.

The present reasoning suggests that at the target performance of an adaptive procedure the ML estimate of the probability should be correct, and this is in line with the findings of Treutwein and Strasburger (1999) that the ML threshold estimates were not biased. However, it has not been proven up to now that ML threshold estimates are more reliable than other ways of calculating the threshold, such as averaging the reversal points or signal levels following an appropriate amount of discard.

Moreover, it may appear inelegant to evaluate the complete form of the PF, knowing it is wrong, in order to extract a single value from it, which might be the only value of the PF estimate that is not biased.

REFERENCES

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., & Brunk, H. D. (1972). *Statistical inference under order restrictions*. Chichester, U.K.: Wiley.

Green, D. M., Richards, V. M., & Forrest, T. G. (1989). Stimulus step size and heterogeneous stimulus conditions in adaptive psychophysics. *Journal of the Acoustical Society of America*, **86**, 629-636.

Kaernbach, C. (1991). Simple adaptive testing with the weighted up–down method. *Perception & Psychophysics*, **49**, 227-229.

King-Smith, P. E., & Rose, D. (1997). Principles of an adaptive method for measuring the slope of the psychometric function. *Vision Research*, **37**, 1595-1604.

Kontsevich, L. L., & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, **39**, 2729-2737.

Leek, M. R., Hanna, T. E., & Marshall, L. (1992). Estimation of psychometric functions from adaptive tracking procedures. *Perception & Psychophysics*, **51**, 247-256.

Levitt, H. (1971). Transformed up–down methods in psychophysics. *Journal of the Acoustical Society of America*, **49**, 467-477.

Miller, J., & Ulrich, R. (2001). On the analysis of psychometric functions: The Spearman–Kärber method. *Perception & Psychophysics*, **63**, 1399-1420.

Pflug, G. C. (1990). Non-asymptotic confidence bounds for stochastic approximation algorithms with constant step size. *Monatshefte für Mathematik*, **110**, 297-314.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, **22**, 400-407.

Strasburger, H. (2001a). Converting between measures of slope of the psychometric function. *Perception & Psychophysics*, **63**, 1348-1355.

Strasburger, H. (2001b). Invariance of the psychometric function for character recognition across the visual field. *Perception & Psychophysics*, **63**, 1356-1376.

Treutwein, B., & Strasburger, H. (1999). Fitting the psychometric function. *Perception & Psychophysics*, **61**, 87-106.

NOTES

1. For the ML estimate of the PF, both threshold and slope parameters were varied. This corresponds to the normal experimental situation in which one does not know the true threshold. In simulations, one knows the true threshold and can make use of this knowledge. ML fits based on the true threshold (i.e., varying only the slope parameter) show much less slope bias (about a quarter) than ML fits where both parameters are varied. If only the slope parameter is varied, one introduces a further kind of slope bias that, being negative, counteracts the positive slope bias from data dependency: Runs that would lead to threshold estimates far from the true threshold will have remarkably low slope estimates when the PF is forced to cross 0.5 at the true threshold (i.e., often outside the region that has been tested).

2. Please note that this effect, although similar to that presented in Note 1, is different in nature and extent. The PFs that enter the averaging are determined by varying both threshold and slope parameters (i.e., no use is made of the knowledge of the true threshold). Whereas for runs with 30 trials (step size 0.5) the slope bias goes down from 39% to 9% when use is made of the knowledge of the true threshold, it goes down to 34% when the PF is averaged. Moreover, although the effect of making use of the knowledge of the true PF persists for all run lengths, the reduction by averaging the PF is only effective for runs of up to 30 trials.

3. Jeff Miller and I are grateful to Stanley Klein for this suggestion. The simulated adaptive runs were prepared in Leipzig, and the ML probit and Spearman–Kärber analyses were done by Jeff Miller.

4. One could think of disentangling these two mechanisms by repeating the analysis that was done for Figure 2B, but this time abstaining from imposing monotonicity and not extrapolating to the untested region. The remaining slope bias could be considered to represent the share of the retest probability, because the second mechanism, the test-at-all probability, cannot be effective when no extrapolation takes place. This estimate of the share of the retest probability increases with increasing run length (10: 40.4%; 20: 54.9%; 30: 62.6%). This is in accordance with intuition: With longer runs, the importance of the test-at-all effect should decrease (as nearly all relevant levels will be tested), and the importance of the retest effect should increase (as retests will take place more frequently). However, it should be noted that averaging without extrapolating implies a selection bias: For levels below the equilibrium point, the averaging will skip those runs in which the level in question was not tested (i.e., those runs that would normally be interpreted as demonstrating a low performance at the level in question). More generally, it might be questioned whether it makes sense to disentangle these two effects that might well be considered two aspects of the same bias source (i.e., serial dependency).

(Manuscript received July 10, 2001;
revision accepted for publication August 8, 2001.)