Using the standard staircase to measure the point of subjective equality: A guide based on computer simulations

T. S. MEESE

University of Birmingham, Edgbaston, Birmingham, England

In experiments that measure a point of subjective equality, it is necessary to employ a psychophysical technique that measures the 50% point (P50) on the psychometric function. The standard staircase is an attractive candidate, because its simplicity makes it easy to understand and implement. However, Pentland (1980, *Perception & Psychophysics*, 28, 377–379) found his own maximum likelihood method (Best PEST) to be considerably less variable than the standard staircase. Here, computer simulations were used to study the effects of manipulating several parameters of the standard staircase in order to try to find a variant that would be competitive with the performance of Best PEST. A graphical summary of some of the standard staircase implementations that were found to be most stable is presented in order that investigators may choose easily the parameters best suited to their requirements. A comparison with an implementation of Pentland's Best PEST failed to replicate Pentland's finding, suggesting that investigators should have no misgivings about the standard staircase in experiments that measure a point of subjective equality.

Adaptive psychophysical procedures are typically used in preference to the method of constant stimuli in order to minimize the duration of data collection when measuring either a point of subjective equality or a sensory threshold (Watson & Fitzhugh, 1990). Adaptive procedures, in which the stimulus placement for each trial is determined by some portion of the response history, can be split into two general categories. First are the staircase procedures, examples of which are the standard and transformed staircases (see, e.g., Cornsweet, 1962; Dixon & Mood, 1948; Levitt, 1971; Wetherill & Levitt, 1965) and PEST (Findlay, 1978; Taylor & Creelman, 1967; Taylor, Forbes, & Creelman, 1983). Essentially, these procedures move up and down the stimulus range in step sizes generated by some simple set of rules. The second category comprises the maximum likelihood procedures. Examples are Best PEST (Pentland, 1980), APE (Watt & Andrews, 1981), QUEST (Laming & Marsh, 1988; Watson & Pelli, 1983), ML-TEST (Harvey, 1986), and the step method (Simpson, 1989). Typically, in order to determine the optimum stimulus placement on each trial, maximum likelihood procedures fit a series of para-

metric models of the psychometric function to the data collected on all previous trials. This same procedure is also used at the end of the session to obtain the best estimate of the point of subjective equality (PSE) or threshold. More details of one such method (Best PEST) are given in the Method section.

Although maximum likelihood procedures tend to be theoretically better motivated than staircase procedures, they require a greater number of assumptions about the form of the psychometric function and can be quite difficult to understand, particularly for researchers without a mathematical background. On the other hand, the standard staircase is easy to understand and implement in computer-controlled experiments (see Method), and so it is often the experimenter's preferred choice. Consequently, it should be of interest to compare the effectiveness of these procedures. Over recent years, Monte Carlo computer simulations have been used by several authors to investigate the relative efficiencies of these procedures (e.g., Emerson, 1984; Findlay, 1978; Green, 1990, 1993; Hall, 1981; Harvey, 1986; Kaernbach, 1990, 1991; King-Smith, Grigsby, Vingrys, Benes, & Supowit, 1994; Leek, Hanna, & Marshall, 1992; Madi-gan & Williams, 1987; Pentland, 1980; Rose, Teller, & Rendleman, 1970; Schlauch & Rose, 1990; Simpson, 1989; Swanson & Birch, 1992; Watson & Fitzhugh, 1990; Watt & Andrews, 1981). In all of these studies, a (reasonable) parametric form of the response probability function (the psychometric function) is assumed in order to model human behavior. The computer model and a random number generator then work through several hundred psychophysical sessions controlled by the adaptive procedure under investigation and generate a

The simulations reported in this paper were performed at the Department of Psychology, University of Bristol; the Department of Vision Sciences, Aston University, Birmingham; and The University of Birmingham, School of Psychology. An extensive summary of the results for the full set of standard staircase conditions tested are available in graphical form by sending a stamped addressed envelope to the author. Much of this work first appeared in abstract form in Meese (1991). Correspondence should be addressed to T. S. Meese, School of Psychology, University of Birmingham, Edgbaston, Birmingham B15 2TT, England.

distribution of estimates of a given target point on the psychometric function. The descriptive statistics of such distributions can then be used to summarize and compare the behavior of different parameter sets, adaptive procedures, and simulated experimental conditions.

Investigators have examined implementations of adaptive procedures that converge on both the 50% point (P50) of the psychometric function and at some point higher than P50. The latter can be used in two-interval forced-choice (2IFC) paradigms, which are typically preferred for detection and discrimination experiments although there are exceptions (see, e.g., Green, 1993, and Kaernbach, 1990). The former are often referred to as yes/no (Y/N) procedures, although they are most usefully employed in matching and nulling experiments in which a PSE is measured.

The concept of a PSE can be usefully illustrated through consideration of the tilt aftereffect (Gibson & Radner, 1937); the finding that adaptation to a grating oriented at, say, 15° can make a vertical test grating appear tilted in the opposite direction. The magnitude of this effect can be measured by using a matching technique and the method of constant stimuli—for example, by comparing the test grating with a range of oriented reference gratings presented in an unadapted region of the visual field. Observers simply report whether the test grating is perceived as being tilted to the left or to the right of each reference grating. The psychometric function is revealed by plotting the percentage of “left” (or “right”) responses as a function of reference orientation, and the PSE is given by the 50% point on this function, determined by a curve-fitting procedure such as probit analysis (see Finney, 1971).

Note that although I shall continue to refer to techniques that measure P50 as Y/N procedures, when they are used to measure a PSE, a task-specific label tends to be more appropriate. For example, in the experiment considered above, the task is better referred to as a left-right (L/R) task, although in the case of a spatial alignment task it has been referred to as one-interval forced choice (1IFC; e.g., Hess & Hayes, 1994). Furthermore, an important note that was emphasized by the referees of the present paper is that most Y/N techniques (but see Green, 1993, and Kaernbach, 1990, for exceptions), including the standard staircase, are unsuitable for the measurement of detection and discrimination thresholds, because they do not take account of false positive responses, which raise the lower limb of the psychometric function (see Viemeister & Schlauch, 1992, for further consideration of this point).

Pentland (1980) is one of the few authors who have investigated the commonly employed Y/N standard staircase. He found that his maximum likelihood algorithm (Best PEST) produced less variable estimates of P50 than did Taylor and Creelman's (1967) PEST, Findlay's (1978) improved PEST, or a standard staircase (Cornsweet, 1962). For the first two methods, this superiority was around a factor of 1.5–2, but for the standard staircase, the potential improvement offered by Best PEST

was about a factor of 10. This is somewhat alarming, given the widespread popularity of the standard staircase.

However, Pentland's (1980) finding is rather surprising in light of 2IFC and 3IFC experiments done with real subjects, in which essentially no differences between PEST (Taylor & Creelman, 1967), Best PEST (Pentland, 1980), QUEST (Watson & Pelli, 1983), and the transformed standard staircase (Wetherill & Levitt, 1965) have been found (Madigan & Williams, 1987; Shelton, Picardi, & Green, 1982; Shelton & Scarrow, 1984). Moreover, Rose et al. (1970) also performed simulations on a Y/N implementation of the standard staircase, using either a normal ogive or a ramp function as a model of the psychometric function. Although Pentland modeled the psychometric function with a logistic function, a rough comparison of the two sets of results shows the performance of Rose et al.'s implementation of the standard staircase to be roughly comparable with that found in Pentland's implementation of Best PEST. Although there are several procedural differences between these two studies, it is not clear why the results differ so greatly. One possible explanation is that Pentland's implementation of the standard staircase was less than optimal. There are several ways in which this could be so. First, Pentland did not include the commonly employed procedure of not including the first few reversals of the staircase in the final analysis. Second, his choice of step size may have been suboptimal. Third, he did not employ Cornsweet's (1962) suggestion of starting the staircase with a large step size and then reducing it to some smaller size after a given number of reversals. This final modification allows the staircase to home in more rapidly on P50.

It was with these ideas in mind that a series of Monte Carlo computer simulations was designed to try to find a variant of the standard staircase that would retain its simplicity and yet achieve a proficiency competitive with that of maximum likelihood methods. A second aim of the study was to explore the robustness of the standard staircase and Best PEST, by measuring the sensitivity of these techniques to simulated lapses in observer attention (Madigan & Williams, 1987; Swanson & Birch, 1992; Taylor, Forbes, & Creelman, 1983; Viemeister & Schlauch, 1992). Finally, a selection of results from over 5.7 million simulated psychophysical trials is summarized in graphical form, to aid parameter selection by investigators who may wish to use the standard staircase in experiments that measure a PSE. This should be particularly useful to workers who are new to psychophysics and have little experience to draw upon when selecting staircase parameters such as step size.

METHOD

The Standard Staircase

Although the logic behind the standard staircase is independent of its application (Levitt, 1971), the algorithm is best explained by way of a concrete example. Consider a tilt aftereffect experiment (see the Introduction) in which the orientations of a test grating and a reference grating are compared. Before commencing data

collection, one must choose both a starting point for the initial orientation of the reference grating and a step size that will determine the magnitude of orientation change between trials. The observer is then presented with both the test and the reference stimuli, and he/she responds by indicating whether the test grating appeared to be oriented to the left or to the right of the reference grating. On the next trial, the orientation of the reference grating is adjusted by an amount equal to the step size and in the direction of the previous response. This procedure is repeated until the direction of the response changes and the stimulus level (in this case, grating orientation) at which this *reversal* occurs is recorded. An example of two such reversals is shown in Figure 1a.

The procedure is continued until a predetermined, even number of reversals have been recorded, and an estimate of P50 is taken as being the mean of the stimulus levels at which the reversals have occurred (Wetherill & Chen, 1964, cited in Wetherill & Levitt, 1965). In order to differentiate this algorithm from the two modifications that follow, it will be referred to as the *single-step algorithm*.

Alternative Standard Staircase Algorithms

In real experiments, observers may on occasion find their attention drifting from the task in hand for a variety of reasons (e.g., tiredness, a sneeze, forgetting the requirements of the task, lack of motivation). Granted the assumption that the observer's response is random if the stimulus is missed because of such lapses, then 50% of these responses will be in the same direction as would most likely have occurred had the stimulus not been missed, whereas the other 50% will be in the opposite direction. This second set of *inappropriate responses* has the most severe consequence on a staircase's estimate of P50 when the stimulus level is distant from P50. This is because the estimate of P50 is based on the mean of the reversal points, and stray reversals that are distant from P50 will obviously lead to bias.

It seems likely that the single-step algorithm will be sensitive to lapses in observer attention because the stimulus level is changed after each response; thus, inappropriate responses are acted upon by the algorithm without question. An attempt to improve on this potential limitation was made by developing two simple extensions to the single-step algorithm. The first of these, referred to as

the *confirmation algorithm*, requires that no change of stimulus level take place until two consecutive responses have been made in the same direction (i.e., two positive responses or two negative responses). Figure 1b shows an example of a pattern of responses with two reversals.

To see the extent to which this algorithm is less sensitive to lapses of attention than the single-step algorithm, consider the following. If we assume that observers respond at chance when their attention lapses, the probability of a single inappropriate response, $p(i)$ is given by

$$p(i) = 0.5\lambda, \quad (1)$$

where λ is the proportion of trials on which lapses of attention occur.

The probability of two consecutive inappropriate responses is therefore given by

$$p(i)^2 = 0.25\lambda^2. \quad (2)$$

Thus, the relative improvement offered by the confirmation algorithm over the single-step algorithm, given by the ratio of Equation 1 to Equation 2, is equal to $2/\lambda$. For example, if the lapse rate is 0.1 (10%), the confirmation algorithm would make only one 20th of the number of inappropriate responses made by the single-step algorithm.

This improved insensitivity to lapses is irrespective of stimulus placement, although, as already mentioned, the consequences of inappropriate responses are most severe when the stimulus level is distant from P50. It is important, however, to realize that this improvement is not without its cost, because for any given number of steps or reversals, the confirmation algorithm will obviously require more trials.

Note that the confirmation algorithm is similar to an implementation of PEST (Taylor & Creelman, 1967) with a target value of 0.5 and $W = 1$, but with a fixed step size (see Taylor & Creelman, 1967). The main differences are that PEST includes a rule for increasing step size, and that PEST uses only the final stimulus placement in order to calculate P50 (but see also Kaplan's RAT mode PEST in Taylor et al., 1983).

The second modified standard staircase algorithm, referred to as the *hybrid algorithm*, is a cross between the single-step algorithm and the confirmation algorithm. In this algorithm, two consecutive alike responses are required only if a change in response direction occurs. An example response pattern for this algorithm is shown in Figure 1c, where two reversals are shown. The intention was to create an algorithm with an insensitivity to lapses of observer attention that was similar to the confirmation algorithm while at the same time requiring fewer trials. For example, during the initial stage of testing, when the stimulus levels may be distant from P50 and the confidence of each response is high, the hybrid algorithm is not delayed by having to confirm every response. Nevertheless, it remains robust against inappropriate responses, because as soon as a potential reversal is indicated, response confirmation is required.

Additional Parameters

In real experiments, it is a common practice to include a few reversals at the beginning of a staircase that are not included in the final calculation of P50. The reasoning is that the first few trials of a session should give observers time to feel comfortable with the task and establish their response criteria. This approach also provides some protection against the influence of inappropriate responses during the early stages of testing when the stimulus level may be distant from P50 and hence potentially more serious, as discussed above. A further modification was suggested by Cornsweet (1962), who recommended that the first few steps of the standard staircase should be set larger than later steps, so that the staircase would home in more rapidly on P50. These two modifi-

Figure 1. A single pair of reversals for three different variants of the standard staircase: (a) the single-step algorithm; (b) the confirmation algorithm; and (c) the hybrid algorithm. The letter N denotes that the response was negative (e.g., rightward) and that the staircase is being driven upward (or leftward), and the letter P denotes that the response was positive (e.g., leftward) and that the staircase is being driven downward (or rightward).

cations can be conveniently implemented by introducing two additional parameters that control a preliminary stage of *reducing reversals*, prior to the main body of the staircase where a series of *effective reversals* is used to calculate P50. The first parameter sets the number of reducing reversals and the second parameter, referred to as the *reduction factor*, determines the rate of decrease in step size after each reducing reversal. More formally, the initial step size (IS), used at the start of testing, is given by

$$\begin{aligned} \text{IS} &= S \cdot \text{RF}^{(n\text{RR}-1)} & \text{for } n\text{RR} > 1, \\ \text{IS} &= S & \text{for } n\text{RR} \leq 1, \end{aligned} \quad (3)$$

where RF is the reduction factor, $n\text{RR}$ is the number of reducing reversals, and S is the minimum step size (i.e., the final step size within the main body of the staircase). For example, if the minimum step size is set to two units, if the reduction factor is two, and if the number of reducing reversals is three, then the initial step size would be eight units. After the first reducing reversal, this would be halved to four units, and after the third reducing reversal, it would be halved again to two units. After the third and final reducing reversal, the step size remains at two units for the remainder of the staircase. Thus, unlike some other methods (see, e.g., Emerson, 1984; Robbins & Monro, 1951; Taylor & Creelman, 1967), the step size is not consistently modified throughout testing. If the stimuli are generated afresh for each presentation (using, e.g., a grating generator), there is no a priori reason for preferring a method with a fixed minimum step size. If, on the other hand, the stimuli are generated prior to experimentation (using, e.g., a frame store), it is necessary to know all of the potential stimuli in advance of testing. With an ever changing step size, the potential number of stimuli can become practically unmanageable in the methods proposed by Robbins and Monro (1951) and Emerson (1984). However, as long as the reduction factor in the standard staircase algorithm is an integer, the potential stimulus set is an array of levels with a fixed interval equal to the minimum step size. Of course this in itself does not provide an upper and lower limit on the array, but in most practical applications, these can be set by the experimenter.

Best PEST

Pentland (1980) dubbed his own maximum likelihood procedure Best PEST, though it bears little resemblance to Taylor and Creelman's (1967) original PEST. Although the computer code presented by Lieberman and Pentland (1982) requires that the stimuli be placed at intervals of 1 logit, this code can easily be modified to allow stimulus spacing to be determined by the user. The stimulus range is also declared by the user in advance of testing and is defined as the difference between the two extremes of the stimulus domain.

Pentland's approach is to calculate the probability that P50 (in a Y/N implementation) lies at each of the available stimulus levels by using the observer's response history to create a probability density function of the variate under investigation. The level with the greatest probability (maximum likelihood) of being P50 is the mode of this distribution and is chosen for the next testing level. In carrying out these calculations, a logistic function (see below) is used as a model of the psychometric function, for which a spread (or slope) parameter must be estimated in advance of testing. The procedure's estimate of the PSE (or threshold) is simply the stimulus level at which the next presentation would be made after a predetermined number of trials.

For a more detailed, but nonmathematical, account of the maximum likelihood procedure, see Green (1993) and Madigan and Williams (1987); and for a mathematical treatment and a computer algorithm of Best PEST written in BASIC, see Lieberman and Pentland (1982). Detailed treatments of an alternative maximum

likelihood algorithm (QUEST) are given by Watson and Pelli (1983), Laming and Marsh (1988), and King-Smith et al. (1994).

The Simulations

As already mentioned, in order to simulate a psychophysical observer, an assumption must be made about the form of the psychometric function. The cumulative Gaussian (normal ogive) is a common choice, and following several others, it was approximated by the logistic function (Emerson, 1984; Findlay, 1978; Madigan & Williams, 1987; Pentland, 1980; Taylor & Creelman, 1967). This function is given by

$$P(+) = 1 / \left(1 + e^{[(P50-L)/E]} \right) \quad (4)$$

where $P(+)$ is the probability of a positive response, L is the stimulus level in logits, E is a parameter that determines the spread (or slope) of the psychometric function, and P50 is the stimulus level for which the probability of a positive response is .5. Note that when a PSE is being measured, $P(+)$ is simply interpreted as the probability of a response in one of the two possible directions (e.g., either left or right). In all of the simulations reported here, E was always set to 1. The reason for this is that although the spread (or slope) of an observer's psychometric function may differ with the task, it is the relation between the spread and the step size that is important. Consequently, it is necessary to manipulate only one of these variables. I choose to manipulate step size, but will also consider the results in the context of an experimenter's confidence in his/her estimation of the spread of the observer's psychometric function in order that the results should have practical significance. Thus, a useful interpretation of step size is to think of it as a measure relative to spread (i.e., in these simulations, relative to 1 logit), rather than as an absolute value. For example, a step size of 0.5 logit should be thought of as a step size of half of the spread (in logits), and a step size of 2 logits should be thought of as a step size of twice the spread (in logits).

Note that some other authors (e.g., Emerson, 1984) have placed E in the numerator of Equation 4 and referred to it as a "slope" parameter. I prefer to place E in the denominator, because this leads to a more natural relation to step size, and I emphasize this by referring to it as a *spread* parameter.

The logistic function was implemented as a simulated observer by comparing the result of Equation 4 with a number drawn pseudorandomly from between zero and one. If the random number was the smaller quantity, the response was deemed to be positive (e.g., left), otherwise it was negative (e.g., right). The random number generator was reseeded by the system clock at the beginning of each run.

In studies of this kind, there have generally been two approaches to choosing where to place the initial stimulus in relation to the target point on the psychometric function. Some authors have assumed that in real experiments the investigator will place the first trial as close as possible to an estimate of the target point (e.g., Madigan & Williams, 1987; Watson & Fitzhugh, 1990), whereas others have assumed that the investigator will present the first stimulus at some distance away from their best estimate (e.g., Emerson, 1984; Findlay, 1978; Pentland, 1980; Taylor & Creelman, 1967). Here, the second approach was adopted and implemented in a standard way as follows. For each simulated run (i.e., each simulated psychophysical session), the starting level was placed at a randomly chosen point along the continuum between 5 and 10 logits above P50. The results are based on the outcome of 500 Monte Carlo simulations.

For a second batch of simulations, the starting level was increased to between 10 and 20 logits above P50.

The standard staircase. Simulations were performed for all three standard staircase algorithms with the smallest step size set

to 0.2, 0.5, 1, and 2 logits; the reduction factor was set to 2, 3, and 4, and the number of reducing reversals was set to 2, 3, and 4. However, a small subset of these parameter values was omitted because of the inappropriately large initial step size that can be produced. For example, if the smallest step size is 2 logits, if the reduction factor is 4, and if there are 4 reducing reversals, the initial step size would be 128 logits. This is far larger than would reasonably be used in practice.

A basic implementation of the standard staircase (basic standard staircase) was also simulated for each algorithm. This implementation had no reducing reversals; the estimate of P50 was based on all reversals and a minimum step size of 1 logit. This is identical to the implementation of the standard staircase reported by Pentland (1980).

Best PEST. The algorithm presented in Lieberman and Pentland (1982) was modified to allow the minimum step size to be adjusted, and it was used to implement Best PEST in order to provide a comparison of a maximum likelihood method with the standard staircase. However, it was not the intention of this paper to find an optimum implementation of Best PEST, but rather to concentrate on the standard staircase. Consequently, the parameter values of Best PEST were not manipulated extensively, but instead were set to values that might reasonably be used in real experiments, and they were based on the results of Emerson (1984) and of Madigan and Williams (1987), as well as on some preliminary simulations (see Results).

Lapses of attention. For the purposes of these simulations, a lapse was considered to be equivalent to an observer's not seeing the stimulus, whereupon it was assumed that the observer would respond at chance (i.e., an even probability of making either response). The simulations were repeated with the lapse rate set to 0%, 15%, and 30% for the full set of parameter values and standard staircase algorithms, as well as for Best PEST.

Bias and variability. The bias of a method is taken as the difference between the true P50 and a procedure's estimate of P50, averaged over 500 Monte Carlo simulations. The variability of a method is taken as the standard deviation of 500 derivations of a procedure's estimate of P50. Note that some authors have referred to this measure as accuracy (e.g., Madigan & Williams, 1987; Pentland, 1980) and precision (e.g., Swanson & Birch, 1992).

The proficiency of a method is conveniently illustrated by plotting both bias and variability against the number of simulated trials. However, for the standard staircase, the estimate of P50 is calculated after each reversal pair rather than after each trial. Consequently, the mean number of trials from the 500 simulations was calculated for each reversal pair and used as the function against which variability and bias were plotted. For Best PEST, the results are shown at five-trial intervals.

For each standard staircase condition tested, the simulated run was terminated after 12 effective reversal pairs.

RESULTS

Preliminary Investigation of Best PEST

Since neither Pentland (1980) nor Lieberman and Pentland (1982) published a full list of the parameter values used in their simulations of Best PEST, these values were derived by other means. The results of some preliminary simulations of Best PEST are shown in Figure 2. The triangular symbols show the results for parameter values set to those published by Madigan and Williams (1987) and tested under their less conservative regime of placing the starting level between ± 3 logits from P50. The results from my own simulations, shown by the filled triangles, and the results replotted from

Figure 2. The variability of different implementations of Best PEST under different simulation regimes for a yes/no task. The open triangles are results replotted from Madigan and Williams (1987), and the filled triangles are the results of Madigan and Williams's simulations replicated here. (Range = 8 logits, spread = 1/5 of the range [1.6 logits], and the starting level is placed randomly ± 3 logits from P50.) The filled circles show the effects of increasing the range to 24 logits and the spread parameter to 3 logits, and randomly placing the starting level between 5 and 10 logits above P50. The filled squares are for the same condition as are the filled circles, but with the spread parameter set to 1 logit. In all cases, the stimulus spacing was 0.5 logit.

Madigan and Williams (1987), shown by the open triangles, are remarkably similar, which helps to validate both the simulations and the calculations of variability reported here.

Madigan and Williams (1987) used an interval of 0.5 logits for the spacing of stimuli in Best PEST, and this spacing was adopted for the remainder of my own simulations of this procedure. However, it was necessary to increase the range parameter of Best PEST from that used by Madigan and Williams in order to accommodate the greater range of potential locations of P50 used in this study. In order to cater for a starting level placed between 5 and 10 logits above P50 (see Method), the range parameter in Best PEST would have to be set to at least 20 logits, because Best PEST always selects its first stimulus from the center of its range. However, this would mean that on some occasions the simulated P50 would fall on the lower extreme of the stimulus range. This would make Best PEST appear slightly less variable than it really is, because good candidate levels for P50 would lie outside the permitted range and so would not be contemplated by the algorithm. It is likely that in real experiments, investigators would aim to avoid this state of affairs. Consequently, the stimulus range was increased to 24 logits.

For a range of 24 logits, the spread parameter was set to both 1 logit (squares in Figure 2) and 3 logits (circles in Figure 2). A spread of 1 logit means that the spread of the estimated psychometric function used by Best PEST is identical to the spread of the simulated observer's psychometric function, whereas a spread of 3 logits means that the estimated spread is somewhat broader than the

observer's spread. Lieberman and Pentland (1982) suggested using a spread set to one fifth of the range and considered it safest to overestimate the breadth of the spread, particularly if lapses of observer attention were expected. Emerson (1984) investigated the effects of spread mismatch in Best PEST and found the procedure to be least variable when the match was perfect, though variability remained good, for mismatches of both 3:1 and 1:3. Because one of the aims of this study was to investigate the effects of lapses of observer attention, a spread parameter of 3 logits was considered to be a fair compromise between optimizing the performance of Best PEST in conditions of both lapsing and no lapsing. For example, when the spread parameter was changed from 1 logit to 3 logits, the variability of Best PEST after 40 trials deteriorated little from 0.36 to 0.46 logit when the lapse rate was 0% (see Figure 2), whereas for a lapse rate of 30%, the variability improved substantially from 2.13 logits to 0.99 logit (not shown). Unless otherwise stated, the spread parameter of Best PEST was set to 3 logits in further simulations.

A Comparison Between the Standard Staircase and Best PEST

For each condition of simulated lapsing, the "optimum" implementation of the standard staircase was identified informally by choosing the algorithm and parameters that produced the lowest variability after 40 simulated trials while retaining a low level of bias.¹ In order to make comparisons after a fixed number of trials, it was necessary to interpolate between the results for reversal pairs that flanked the 40 trial point. Linear interpolation was considered to be an adequate procedure for this task.

The parameter sets and algorithms for the optimum implementations of the standard staircase are summarized in the top three rows of Table 1. Detailed results are shown in Figure 3, where variability (left column) and bias (right column) are shown for the optimum standard staircase (filled diamonds), the basic standard staircase (filled squares), and Best PEST (open circles) for each condition of simulated lapsing (different rows). The first noteworthy point is illustrated in Figure 3a, where the basic standard staircase is shown to be comparable to Best PEST. This is surprising, because the conditions for the basic standard staircase are exactly the same as those published by Pentland (1980), who found the standard

staircase to have a variability of approximately 4 logits after 40 trials. The simulations performed here show the basic standard staircase to be about 10 times less variable than that found by Pentland (1980).

The second point to note is that for all three conditions of simulated lapsing, the optimum standard staircase is either superior or comparable to the basic standard staircase and Best PEST. The improvements offered by the optimum standard staircase over the basic standard staircase are particularly noticeable in terms of bias, especially under the conditions where the rate of simulated lapsing was greater than 0% (see Figures 3d and 3f). Both Best PEST and the optimum standard staircase show good to tolerable performances, even under the condition of 30% simulated lapsing, suggesting that either of these techniques may be useful when lapses in an observer's attention are foreseen to be a potential problem.

Further Comparisons With the Standard Staircase

Figure 4 shows variability for the optimum standard staircase alongside alternative techniques replotted from Findlay (1978), Pentland (1980), and Emerson (1984). The results for Taylor and Creelman's (1967) PEST are taken from Findlay (1978). (Compare Figure 3a and Figure 4 with Pentland's [1980] Figure 1 to see how much better the standard staircase has fared in the simulations reported here.) For each set of results shown in Figure 4, the simulation regime was identical to that used here, allowing meaningful comparisons to be made across different papers. The optimum standard staircase is superior or comparable to all of the other techniques, apart from Pentland's (1980) implementation of Best PEST, for which he does not report the parameter values.

Increased Investigator Uncertainty

One concern with the standard staircase results is that the optimum parameter choices may be particular to the range and distance between the starting level and the simulated P50 that were employed. Because it is possible that the investigator may have less idea about the location of the observer's PSE than has been hitherto assumed, the whole series of simulations was repeated with the starting level placed between 10 and 20 logits above P50. This range for the starting level is both broader and more distant from P50 than that used in the previous batch of simulations. To simulate Best PEST

Table 1
Optimum Standard Staircase Algorithms and Parameter Sets After 40 Trials

Start Distance From P50 (Logits)	Lapse Rate (%)	Algorithm	Minimum Step Size	Reduction Factor	Number of Reducing Reversals	Number of Effective Reversals	Variability (Logits)	Bias (Logits)
5-10	0	Single step	0.5	3	4	17.6	0.36	-0.04
5-10	15	Single step	0.5	2	4	16.5	0.48	0.01
5-10	30	Hybrid	0.5	2	3	3.9	0.68	0.09
10-20	0	Single step	1	4	2	19.5	0.38	-0.06
10-20	15	Hybrid	0.2	3	4	2.7	0.51	-0.09
10-20	30	Hybrid	1	2	3	4.8	0.77	0.16

Figure 3. a, c, and e: Variability as a function of trial number for Best PEST, the basic standard staircase, and the optimum standard staircase at lapse rates of 0%, 15%, and 30%, respectively. b, d, and f: Bias as a function of trial number for the same conditions, algorithms, and parameters as those shown in a, c, and e. In all cases, the starting level was placed 5–10 logits above P50.

under these conditions, its range parameter was increased to 44 logits and simulations were performed with the spread parameter set to 3 logits (as before) and 6 logits.

The optimum algorithms and parameter values found for the standard staircase under this starting condition are shown in the bottom three rows of Table 1.

Figure 5 shows the variability (left column) and bias (right column) for the optimum standard staircase (filled triangles), the basic standard staircase (filled squares), and Best PEST (open symbols) for each rate of simulated lapsing (different rows). Also shown are standard staircase results for the algorithm and parameter sets shown in the top three lines of Table 1, but under the new simulation regime (filled diamonds). Once again the optimum standard staircase is comparable or superior to

the performance of both Best PEST and the basic standard staircase in terms of both variability and bias. Indeed, in the previous instances in which the standard staircase showed superior performance over its rivals, this margin can be seen to have increased (compare the filled triangles with the filled squares and open circles in Figure 5).

In this batch of simulations, the most dramatic observation is the superiority of the optimum standard staircase over the basic standard staircase in the 15% and 30% lapse rate conditions. These results also show that for the case of 0% lapses, the optimum algorithm and parameters found in the previous batch of simulations (i.e., those in the top three lines of Table 1) are not critically dependent on the position of the initial stimulus in rela-

Figure 4. Comparison of the variabilities of different psychophysical procedures intended to measure P50. All procedures used a logistic function with a spread parameter of 1 logit as a model of the psychometric function. In all cases, the first stimulus was placed at a level selected randomly from between 5 and 10 logits above the simulated P50. The filled symbols show the results for the optimum standard staircase reported here, and the open symbols are results replotted from the authors shown in the key. The PEST results (Taylor & Creelman, 1967) are the standard implementation of PEST with the W parameter of the Wald test set to 1 and are replotted from Findlay (1978).

tion to P50. This is illustrated by the similarity between the results for the two different sets of optimum standard staircase parameters in Figure 5 (compare the filled triangles and filled diamonds in Figures 5a and 5b).² However, for lapse rates of 15% and 30%, performance deteriorates a little if the algorithm and parameters are chosen from the top three lines of Table 1 (compare the filled triangles and filled diamonds in Figures 5c–5f).

This batch of simulations is also particularly revealing of some of the limitations of Best PEST. When the spread parameter was set to 3 logits (as before) and the lapse rate was 30%, the procedure was found to be both biased and extremely variable (open circles in Figures 5e and 5f). However, when the spread parameter was doubled to 6 logits, the performance of Best PEST improved greatly (compare the open triangles and open circles), though it did not achieve that found for the optimum standard staircase. In further simulations (results not shown), the spread parameter was increased to 9 logits so as to be approximately one fifth of the stimulus range, as suggested by Lieberman and Pentland (1982). A comparison of these results with those for a spread parameter of 6 logits (open triangles in Figure 5) revealed a marginal improvement in both variability and bias at a lapse rate of 30% in favor of the broader spread (though it remained inferior to the standard staircase), but for lapse rates of 0% and 15%, there was a marginal deterioration in performance.

Asymmetric Stimulus Placement

One limitation of the simulation regime used here (and that used by many others) is that it is insensitive to one particular potential source of bias. This can be real-

ized by considering the following. In a real experiment, it may be the case that there is only enough time to run one staircase session (e.g., 40 trials) on a particular subject. In advance of testing, the investigator does not know the precise location of P50 on the observer's psychometric function. This means that once the staircase has reached its minimum step size and has approached P50, the stimulus levels that flank P50 are most likely to be asymmetrically placed. For example, if the minimum step size is 8 units and P50 is at 20 units, then, by chance, the flanking stimuli may be placed at, say, 18 units and 26 units. That there is potential for bias here is obvious if it is assumed that the spread of the observer's psychometric function is so narrow that the observer always responds "yes" when the stimulus is above P50 and "no" when the stimulus is below P50. If this were so, a reversal would occur after each response, and the mean of these reversals would give an estimated P50 of 22 units, 2 units above the true P50.

It is important to know whether the asymmetry discussed above will lead to measurement bias when one uses the step sizes and short run lengths (e.g., 40 trials) that are typical in psychophysical experiments. Unfortunately, the simulations that have been presented so far are insensitive to this potential source of bias because the starting level was selected randomly from a stimulus continuum. This means that the occurrence of an asymmetry in one direction is as likely as that in the other direction, and, over the 500 simulated sessions, any bias due to asymmetric stimulus placement would be canceled out.

In order to investigate this issue, Rose et al. (1970) selected the initial stimulus from a discrete set of stimuli whose spacing was the same as the (minimum) step size. The placing of these stimulus levels relative to the simulated P50 was fixed for each block of Monte Carlo simulations, but was adjusted between blocks so that symmetric and asymmetric stimulus placement could be compared. Rose et al. used a ramp function to simulate the psychometric function, and they found essentially no effect for asymmetric stimulus placement. Furthermore, I have confirmed their finding, using a step size of 2 logits and the psychometric function given in Equation 4 (results not shown). For an even larger step size of 4 logits, the worst case of bias found was that for an asymmetry of +1 and -3 logits relative to P50, where after 12 effective reversal pairs the measured bias was -0.204 logit and the variability was 0.528 logit (results not shown). In Monte Carlo simulations at least, the effect of asymmetric stimulus placement appears to be of no practical concern.

Overview of the Standard Staircase Results

The need for an overview. Figures 6–8 and the following remarks were produced as a guide for selecting standard staircase algorithms and parameter sets that best suit the potential user. There were two main reasons for doing this. The first follows from the fact that in order to set the minimum step size according to those

Figure 5. a, c, and e: Variability as a function of trial number for Best PEST, the basic standard staircase, and the optimum standard staircase at lapse rates of 0%, 15%, and 30%, respectively. b, d, and f: Bias as a function of trial number for the same conditions, algorithms, and parameters as those shown in a, c, and e. In all cases, the initial stimulus level was placed 10–20 logits above P50.

shown in Table 1, it would be necessary for the investigator to know the spread of the “true” psychometric function under investigation. For example, suppose an investigator wishes to set the minimum step size to 0.5 logit. Recall, from earlier, that this step size should be viewed as relative to the spread of the psychometric function. Since the spread of the simulated psychometric function was 1 logit, a step size of 0.5 logit corresponds to half of the spread. The investigator must now convert this fraction into units of the experimental variable by multiplying it by the best estimate of the spread of the observer’s psychometric function (i.e., E in Equation 4). A measure of spread more common than E is σ —the standard deviation of a normal ogive—

which approximates E if it is multiplied by 0.6 (Taylor & Creelman, 1967). For example, let us suppose that the standard deviation (σ) of the normal ogive (assumed psychometric function) is estimated at 10 stimulus units. Converting this to E gives 6 stimulus units. This value is then multiplied by the required step size (0.5 in this example) to give a minimum step size of 3 stimulus units.

However, investigators cannot know the spread of the psychometric function exactly; rather, they must make some rough estimation. It is important, then, to determine whether the usefulness of the standard staircase is critically dependent on the accuracy with which the spread of the psychometric function can be estimated.

The second reason for presenting the following overview is that an investigator may wish to use a single algorithm and single set of parameters for a potentially wide range of circumstances. For example, the investigator may suspect that an observer's lapse rate is quite high, but on the other hand it may also be quite low. It may also be the case that on some occasions P50 may be fairly close to the starting level, whereas on other occasions it may be rather distant. In other words, it is helpful to know whether there are a single algorithm and parameter set that perform well under both high and low lapse rates, as well as starting levels that are both fairly close and fairly distant from P50. Table 1 clearly does not provide this kind of information.

Selection of useful and representative data. The complete set of results for the standard staircase algorithms is too extensive to be presented here in its entirety, so that to permit the following overview it was

necessary to perform some judicious pruning of the results. In what follows, all comparisons are made at the 40-trial mark, because typically, performance improved only very slowly as the number of trials increased beyond this point.

Although under some conditions the standard staircase was found to perform best with a reduction factor of 3 or 4 (see Table 1), the improvement over using a reduction factor of 2 was generally found to be marginal. Furthermore, when a reduction factor greater than 2 was used, some parameter sets resulted in unreasonably large initial step sizes (see Method). Therefore, in the following comparisons, only a reduction factor of 2 is considered. Finally, although the confirmation algorithm performed quite well with some parameter sets, the performance of either the single-step algorithm (when the lapse rate was low) or the hybrid algorithm (regardless of lapse rate) was generally comparable or superior. Consequently, the confirmation algorithm appears to

Figure 6. Variability of the standard staircase after 40 trials, using a reduction factor of 2 for three different lapse rates. The top axis indicates the minimum step size, and the bottom axis indicates the number of reducing reversals. Results are shown for the single-step algorithm (a and c) and the hybrid algorithm (b and d). For (a) and (b), the initial stimulus was placed between 5 and 10 logits above the simulated P50, and in (c) and (d), the initial stimulus was placed between 10 and 20 logits above the simulated P50. The asterisks indicate that two effective reversals had not taken place after 40 trials, so that no summary data are available.

have little practical value (given that superior alternatives exist), and it will not be discussed further.

Variability. Figure 6 shows variability (left ordinate) after 40 trials for the single-step algorithm (left column) and the hybrid algorithm (right column) when the starting level was placed between 5 and 10 logits above P50 (top row) and between 10 and 20 logits above P50 (bottom row). The top abscissa indicates the minimum step size in logits and the bottom abscissa indicates the number of reducing reversals. Lapse rate is indicated by the three different styles of shading. The asterisks indicate that a single pair of reversals had not taken place after 40 trials, so no summary data are available for the corresponding parameter sets.

Let us first consider the summary results for a lapse rate of 0%. It is clear from these results that the standard staircase is effected little by step size over the range tested and consequently is not critically dependent on an accurate estimation of the spread of the psychometric function as discussed above. For example, a minimum step size that is twice the spread of the psychometric

function (2 logits in Figure 6) is practically as stable as a minimum step size that is one fifth of the spread of the psychometric function (0.2 logits in Figure 6). Thus, if a minimum step size is chosen from approximately the middle of this range (e.g., about 0.65 of E), the true spread of the psychometric function under investigation could be either three times broader or three times narrower than that estimated, with practically no effect on the performance of the staircase. This is true, regardless of whether the starting level was placed between 5 and 10 logits above P50 or between 10 and 20 logits above P50.

However, for lapse rates of 15% and 30%, it is clear that the hybrid algorithm is generally more stable than the single-step algorithm. For the single-step algorithm, a large minimum step size appears to be more robust against high lapse rates than a small minimum step size, and the variability is perhaps tolerable if the starting level is not too distant from P50 (see Figure 6a). On the other hand, the hybrid algorithm is less sensitive to step size in conditions of simulated lapsing; though if the starting

Figure 7. The same as Figure 6, only for bias.

level may be distant from P50, a choice of 4 reducing reversals and a preference for larger rather than smaller step sizes would appear to be most wise (see Figure 6d).

Bias. Figure 7 shows bias for the same algorithms and parameter sets as those shown in Figure 6 and should be interpreted in the same way. It is clear that both algorithms are essentially bias free, though a small amount of bias can be seen for the single-step algorithm for some parameter sets. This could be avoided by selecting either a large step size, or—for example, if there is a large uncertainty over the spread of the observer's psychometric function—by using four reducing reversals in preference to two or three.

However, for lapse rates greater than 0%, the single-step algorithm can become seriously biased, particularly if the starting level is distant from P50 (see Figure 7c). This is considerably less the case for the hybrid algorithm, where a choice of four reducing reversals provides the best protection against bias (see Figure 7d), though this is not essential if the starting level is known not to be distant from P50 (Figure 7b).

In general, a selection policy that chooses a standard staircase algorithm and parameter set on the basis of the bias results would make recommendations similar to those based on the variability results. Practically, this is convenient, because it avoids the problem of having to trade off bias against variability.

Number of effective reversals. Figure 8 shows the number of effective reversals for the same algorithms and parameter sets as those shown in Figures 6 and 7 and should be interpreted in the same way.

The number of effective reversals required for testing to be expected to terminate after about 40 trials is an increasing function of step size for both algorithms. Not surprisingly, for any given parameter set, the hybrid algorithm encountered fewer reversals than did the single-step algorithm after 40 trials (see Figure 1 and compare the left and right columns of Figure 8).

It is important to remember that all of the results shown in Figure 8 were derived from (linearly) interpolating between flanking data where an even number of reversals had taken place. In other words, if the number

Figure 8. The same as Figure 6, only for number of effective reversals.

of effective reversals is selected from Figure 8 for a real experiment, care should be taken to round (up) to the nearest reversal *pair*, to avoid unnecessary introduction of bias. A second point of practical concern is that, as already mentioned, the required number of effective reversals is dependent on step size. However, if the investigator is uncertain about the spread of the psychometric function, there must also be uncertainty about the step size when it is expressed in terms of the spread parameter. (Remember, manipulating step size is equivalent to manipulating the spread of the psychometric function, and accordingly, uncertainty in one domain is equivalent to uncertainty in the other domain.) This problem can be overcome with a simple modification of the implementation of the standard staircase. Rather than terminating data collection after a predetermined number of effective reversals, termination could be made after some criterion number of trials (e.g., 40 for Figure 6). However, this modification should only be implemented with the proviso that the number of effective reversals that have been made is *even*. In other words, if after the criterion number of trials, the number of reversals that have been made is odd, then data collection should continue until the next reversal; otherwise, data collection should terminate immediately.

DISCUSSION

Best PEST

One of the aims of this study was to investigate the robustness of both the standard staircase and Best PEST to lapses of observer attention. In agreement with the recommendation of Lieberman and Pentland (1982), Best PEST was found to be more robust to lapses when the spread of the psychometric function was overestimated. An insight into why this might be so can be achieved when lapses of attention are treated as a transformation of the psychometric function in order to derive the simulated observer's *effective* psychometric function (see, e.g., Green, 1990). In order to do this, Equation 4 becomes

$$P(+) = \lambda/2 + (1 - \lambda) / (1 + e^{[(P50 - L)/E]}), \quad (5)$$

where λ is the proportion of trials on which lapses of attention occur. Thus, the lower limb of the observer's effective psychometric function asymptotes at $\lambda/2$ and the upper limb asymptotes at $1 - \lambda/2$. This is shown by the filled curve in Figure 9 for a lapse rate of 30% ($\lambda = 0.3$) and the spread parameter (E) set to 1. The dashed curves show λ equal to zero (i.e., equivalent to Equation 4) and spread parameters set to 1 (coarsely dashed curve) and 3 (finely dashed curve). These dashed curves represent two possible forms of the psychometric function used by Best PEST in the simulations reported here. Note that the range of stimulus levels and the placing of P50 have been arranged to mimic a typical Best PEST condition from the first batch of simulations, in which the starting

Figure 9. Comparison between the simulated observer's effective psychometric function after adjusting for a lapse rate of 30% (filled curve), with two forms of estimated psychometric function used by Best PEST (dashed curves). P50 is placed 7.5 logits below the location of the initial stimulus (0 logit) and is typical of the Best PEST simulations for which the results are shown in Figure 3.

level was placed between 5 and 10 logits above P50 (see Figures 2 and 3). Figure 9 shows that neither of the dashed curves is a good approximation to the observer's psychometric function when the lapse rate is 30% (filled curve). However, Best PEST was found to perform considerably better when the function with the broader spread (shallower slope) was used, suggesting that a broad spread better approximates the observer's psychometric function under conditions of lapsing. This is shown more clearly in Figure 10, where the squared difference between the observer's effective psychometric function and the Best PEST functions shown in Figure 9 are plotted. This provides a convenient illustration of the mismatch between the observer's effective psychometric function and those used by Best PEST. Note that P50 is placed at -7.5 logits, and that around this region the squared differences are similar for the two functions. However, farther away from P50, the mismatch is considerably less for the Best PEST function with the broader spread and so confirms the speculation above.

An alternative way to deal with lapses in observer attention in Best PEST is to introduce a user-defined parameter that sets the lapse rate in the psychometric function used by Best PEST. In other words, if a lapse rate of 30% is expected, then Best PEST could be adjusted to use the form of the psychometric function illustrated by the filled curve in Figure 9. However, further simulations will be required in order to investigate whether the potential improvement is justified and to explore how well Best PEST performs in using a psychometric function that is adjusted for lapses but when the observer remains fully attentive.

The Standard Staircase

One of the main advantages of the standard staircase is that few assumptions are required about the form of the observer's psychometric function. The simulations re-

Figure 10. Mismatch, expressed as squared difference, between the simulated observer's effective psychometric function (filled curve in Figure 9) and two estimations of the psychometric function used by Best PEST (dashed curves in Figure 9). The Best PEST function that has the overall best match with the observer's effective psychometric function is the curve that bounds the smallest area with the axes. Thus, the function with a spread parameter of 3 logits (finely dashed curve), is a better match to the effective psychometric function than is the function with a spread parameter that is identical to that of the effective psychometric function (coarsely dashed curve).

ported here show that this is true with regard to both spread and rate of lapses in observer attention. Indeed, the hybrid algorithm was designed in an attempt to produce a staircase algorithm that would meet the latter need, and the results shown in Figures 6 and 7 indicate that the attempt was fairly successful. It should be noted, however, that the single-step algorithm is also remarkably robust against lapses of attention for certain parameter sets.

The results also highlight a second major benefit of the standard staircase. It is often the case that an investigator is not merely concerned with an observer's responses for one particular condition but wishes to learn something about a response characteristic resulting from the manipulation of an experimental parameter. In such cases, if the range of P50 is large and if the proficiency of the psychophysical technique deteriorates when the starting level is distant from P50, then different starting levels must be calculated for each value of the manipulated parameter. This has the potential of being time consuming and error prone, especially if there are individual differences between observers. However, the standard staircase could be used with a fixed starting level for all conditions, since it has been shown to perform well for a wide range of distances between the initial stimulus level and P50.

CONCLUSIONS

The high variability of the standard staircase found by Pentland (1980) has not been replicated here; instead, the findings of Rose et al. (1970) have been supported.

The results of the simulations suggest that investigators wishing to run experiments that measure P50 should have no misgivings about the single-step algorithm. The

simplicity of this algorithm and its comparative performance with Best PEST make it an attractive choice. Indeed, for a 0% lapse rate, the finding that the standard staircase is insensitive to step size over the range tested is a powerful benefit, because this means that the procedure is robust against errors in estimating the spread of the observer's psychometric function.

Finally, in cases in which lapses in observer attention may be foreseen as a problem (e.g., clinical and infant studies), the hybrid algorithm would appear to be a good alternative choice to the single-step algorithm or Best PEST.

REFERENCES

- CORNSWEET, T. N. (1962). The staircase-method in psychophysics. *American Journal of Psychology*, **75**, 485-491.
- DIXON, W. J., & MOOD, A. M. (1948). A method for obtaining and analyzing sensitivity data. *Journal of the American Statistical Association*, **43**, 109-126.
- EMERSON, P. L. (1984). Observations on a maximum likelihood method of sequential threshold estimation and a simplified approximation. *Perception & Psychophysics*, **36**, 199-203.
- FINDLAY, J. M. (1978). Estimate on probability functions: A more virulent PEST. *Perception & Psychophysics*, **23**, 181-185.
- FINNEY, D. J. (1971). *Probit analysis*. Cambridge: Cambridge University Press.
- GIBSON, J. J., & RADNER, M. (1937). Adaptation, aftereffect and contrast in the perception of tilted lines: I. Quantitative studies. *Journal of Experimental Psychology*, **20**, 453-467.
- GREEN, D. M. (1990). Stimulus selection in adaptive psychophysical procedures. *Journal of the Acoustical Society of America*, **87**, 2662-2674.
- GREEN, D. M. (1993). A maximum-likelihood method for estimating thresholds in a yes-no task. *Journal of the Acoustical Society of America*, **93**, 2096-2105.
- HALL, J. L. (1981). Hybrid adaptive procedure for estimation of psychometric functions. *Journal of the Acoustical Society of America*, **69**, 1763-1769.
- HARVEY, L. O., JR. (1986). Efficient estimation of sensory thresholds. *Behavior Research Methods, Instruments, & Computers*, **18**, 623-632.
- HESS, R. F., & HAYES, A. (1994). The coding of spatial position by the human visual system: Effects of spatial scale and retinal eccentricity. *Vision Research*, **34**, 625-643.
- KAERNBACH, C. (1990). A single-interval adjustment-matrix (SIAM) procedure for unbiased adaptive testing. *Journal of the Acoustical Society of America*, **88**, 2645-2655.
- KAERNBACH, C. (1991). Simple adaptive testing with the weighted up-down method. *Perception & Psychophysics*, **49**, 227-229.
- KING-SMITH, P. E., GRIGSBY, S. S., VINGRYS, A. J., BENES, S. C., & SUPOWIT, A. (1994). Efficient and unbiased modifications of the QUEST threshold method: Theory, simulations, experimental evaluation and practical implementation. *Vision Research*, **34**, 885-912.
- LAMING, D., & MARSH, D. (1988). Some performance tests of QUEST on measurements of vibrotactile thresholds. *Perception & Psychophysics*, **44**, 99-107.
- 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 psychoacoustics. *Journal of the Acoustical Society of America*, **49**, 467-477.
- LIEBERMAN, H. R., & PENTLAND, A. P. (1982). Microcomputer-based estimation of psychophysical thresholds: The Best PEST. *Behavior Research Methods & Instrumentation*, **14**, 21-25.
- MADIGAN, R., & WILLIAMS, D. (1987). Maximum-likelihood psychometric procedures in two-alternative forced-choice: Evaluation and recommendations. *Perception & Psychophysics*, **42**, 240-249.

- MEESE, T. S. (1991). The staircase method revisited: A computer simulation. *Perception*, **20**, 113.
- PENTLAND, A. [P.] (1980). Maximum likelihood estimation: The best PEST. *Perception & Psychophysics*, **28**, 377-379.
- ROBBINS, H., & MONRO, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, **22**, 400-407.
- ROSE, R. M., TELLER, D. Y., & RENDLEMAN, P. (1970). Statistical properties of staircase estimates. *Perception & Psychophysics*, **8**, 199-204.
- SCHLAUCH, R. S., & ROSE, R. M. (1990). Two-, three-, and four-interval forced-choice staircase procedures: Estimator bias and efficiency. *Journal of the Acoustical Society of America*, **88**, 732-740.
- SHELTON, B. R., PICARDI, M. C., & GREEN, D. M. (1982). Comparison of three adaptive psychophysical procedures. *Journal of the Acoustical Society of America*, **71**, 1527-1533.
- SHELTON, B. R., & SCARROW, I. (1984). Two-alternative versus three-alternative procedures for threshold estimation. *Perception & Psychophysics*, **35**, 385-392.
- SIMPSON, W. A. (1989). The step method: A new adaptive psychophysical procedure. *Perception & Psychophysics*, **45**, 572-576.
- SWANSON, W. H., & BIRCH, E. E. (1992). Extracting thresholds from noisy psychophysical data. *Perception & Psychophysics*, **51**, 409-422.
- TAYLOR, M. M., & CREELMAN, C. D. (1967). PEST: Efficient estimates on probability functions. *Journal of the Acoustical Society of America*, **41**, 782-787.
- TAYLOR, M. M., FORBES, S. M., & CREELMAN, C. D. (1983). PEST reduces bias in forced choice psychophysics. *Journal of the Acoustical Society of America*, **74**, 1367-1374.
- VIEMEISTER, N. F., & SCHLAUCH, R. S. (1992). Issues in infant psychoacoustics. In L. A. Werner & E. W. Rubel (Eds.), *Developmental psychoacoustics* (pp. 191-209). Washington, DC: American Psychological Association.
- WATSON, A. B., & FITZHUGH, A. (1990). The method of constant stimuli is inefficient. *Perception & Psychophysics*, **47**, 87-91.
- WATSON, A. B., & PELLI, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, **33**, 113-120.
- WATT, R. J., & ANDREWS, P. (1981). APE: Adaptive probit estimation of psychometric functions. *Current Psychological Reviews*, **1**, 205-214.
- WETHERILL, G. B., & LEVITT, H. (1965). Sequential estimation of points on a psychometric function. *British Journal of Mathematical & Statistical Psychology*, **18**, 1-10.

NOTES

1. The differences between many of the implementations were extremely small. Consequently, the term optimum standard staircase is best thought of as an example of one of several proficient implementations of the standard staircase.

2. Close inspection of Figure 5a reveals that the performance of the optimum standard staircase for the 10-20 logit start condition is slightly superior when the optimum parameters found for the 5-10 logit start condition are used. This seemingly nonsensical state of affairs was able to occur because the simulation results in the figures are reruns of the original simulations from which the optimum standard staircase parameters were selected (i.e., one complete set of simulations was performed in order to compare variabilities after 40 trials and determine the optimum parameter set. A second set of simulations was then performed, using these optimum parameters in order to collect full data sets for Figures 3 and 5). Because Monte Carlo simulations are probabilistic, it is quite possible for a rerun simulation to produce slightly different results. This difference should be small, but in this case, when the starting level was placed between 10 and 20 logits above P50, the difference between the original simulations of the algorithms and parameter sets shown in lines 1 and 4 of Table 1 was even smaller.

(Manuscript received November 15, 1993;
revision accepted for publication October 3, 1994.)