Perception & Psychophysics
2001, 63 (8), 1279-1292

Adaptive procedures in psychophysical research

MARJORIE R. LEEK
Walter Reed Army Medical Center, Washington, D. C.

As research on sensation and perception has grown more sophisticated during the last century, new adaptive methodologies have been developed to increase efficiency and reliability of measurement. An experimental procedure is said to be adaptive if the physical characteristics of the stimuli on each trial are determined by the stimuli and responses that occurred in the previous trial or sequence of trials. In this paper, the general development of adaptive procedures is described, and three commonly used methods are reviewed. Typically, a threshold value is measured using these methods, and, in some cases, other characteristics of the psychometric function underlying perceptual performance, such as slope, may be developed. Results of simulations and experiments with human subjects are reviewed to evaluate the utility of these adaptive procedures and the special circumstances under which one might be superior to another.

The study of sensation and perception in humans and other animals is expensive and involves a number of pitfalls and difficulties. Fechner, in 1860, first recognized that inner consciousness might be measurable by outward behavior. It is the measurement of that behavior directly, interpreted as an indirect measure of perception, that is the purview of psychophysics. Over more than a century, methods have been developed and refined that support the systematic exploration within sensory systems of the limits of detection and discrimination among similar and confusable physical stimuli.

The measurement methodologies developed since Fechner's realization have as their primary goal the valid reflection of sensory events. In order to have confidence in the validity and reliability of such measures, many samples of a given behavior typically must be observed in a structured and systematic process, in response to carefully constructed stimuli. Changes in stimulus strength or other characteristics are associated with changes in the ability to detect or discriminate such stimuli. Measures of performance on psychophysical tasks as a function of stimulus strength or other characteristics constitute a psychometric function. A complete characterization of psychometric performance as a function of changes in stimulus strength may be developed by using the method of constant stimuli, one of the classical psychophysical methodologies. With this procedure, a set of stimuli with strengths spanning the range of sensation from imperceptible to consistently perceptible is created. Each member of the stimulus set is presented to an observer many times, at random, and an observation response is requested after each presentation. The psychometric function may be sampled by evaluating the percentage of presentations of each member of the stimulus set that is detected. The function is assumed to be continuous along the stimulus axis, usually with monotonic increases in performance being associated with increasing stimulus strength.

Figure 1 shows an example of a psychometric function, measuring the number of times a particular auditory stimulus was heard, depending on what the strength of the stimulus was. In this example, a method of determining the response to a given stimulus was used that is called a *yes–no* method: On each stimulus presentation, the observer gives one of those two responses, indicating whether the stimulus was perceived or not. Other methods have been developed to make this determination, such as those that involve two or more sequential presentations that differ along some characteristic of interest. The observer is asked to indicate which of the multiple sequential presentations on each trial was a target stimulus. Because the target is always present in one and only one of the presentations on a trial and the observer must select one of the presentations as a response, these are called *forced-choice* methods. Perhaps the most common of the forced-choice methods is the two-alternative forced choice (2AFC), although as will be shown later, forced-choice procedures with three or four alternatives provide more satisfactory measurement of psychometric performance.

Although such a function is assumed to underlie the perception of sensory stimuli, often only one or two parameters of the function will suffice to summarize perception. The most commonly determined parameter is a measure of location of the function along the stimulus axis, typically specified as a threshold stimulus value. The threshold is determined as a level of detection (or discrimination) performance, and frequently the criterion performance for threshold is selected to be the midpoint of

This work was supported by Grant DC 00626 from the National Institutes of Health. The opinions or assertions contained herein are the private views of the author and are not to be construed as official or as reflecting the views of the Department of the Army or the Department of Defense. Correspondence concerning this article should be addressed to M. R. Leek, Army Audiology and Speech Center, Walter Reed Army Medical Center, 6900 Georgia Ave., NW, Washington, DC 20307-5001 (e-mail: leekmar@aol.com).

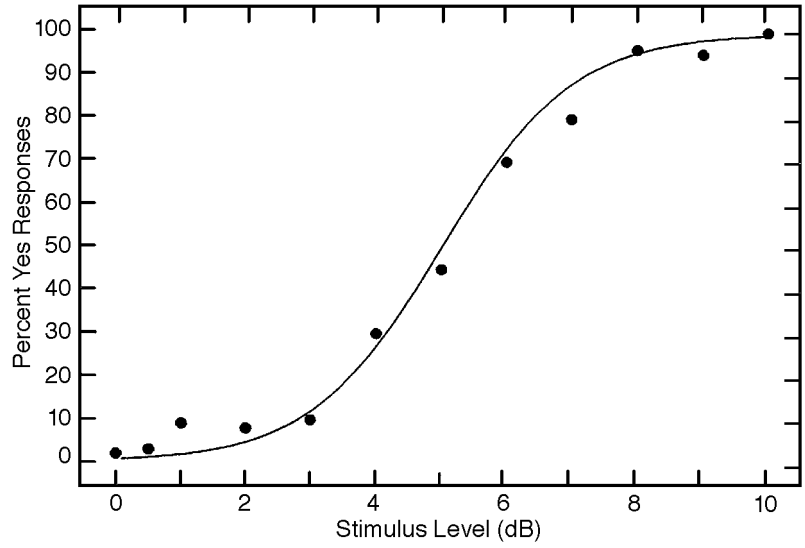

Figure 1. Example of a psychometric function showing percentage of correct detections as a function of stimulus level. The individual data points are fitted with a logistic psychometric function.

the function spanning the range from chance performance to perfect performance. A second summary parameter used to describe performance is the slope of the psychometric function, which is a measure of how rapidly performance changes with a given change in stimulus value. Often, sensory capabilities can be adequately described by a threshold measure alone—that is, one single point on the psychometric function. However, in the method of constant stimuli, the threshold is extracted from a fully sampled function, making the measurement of this single point on the function very expensive in terms of experiment time. Of necessity, many trials are placed at stimulus levels of the underlying psychometric function that are not informative about threshold. It is necessary to have information about performance at these off-threshold levels in order to fully develop the psychometric function, but in many cases that additional information is neither necessary to the goals of the experiment nor worth the extra experimental time and effort. Adaptive psychometric procedures have been developed to address this major problem in psychophysical measurement—that is, an inefficient placement of trials along the stimulus array in order to extract a relevant measure. An experimental procedure is adaptive if the placement of each trial along the stimulus array is determined by the results of the trial or trials that have gone before. It is a characteristic of all adaptive procedures that knowledge about the outcome (e.g., a threshold) increases systematically as the procedure is in progress. That is, the selection of stimuli is determined during the course of the experiment, and stimulus placement is driven by the adaptive algorithm toward the desired measurement point.

Adaptive procedures are designed to rapidly extract relevant measurements from a psychometric function thought to underlie performance on a given sensory/perceptual task. Generally, two types of measures are of interest: location along a stimulus axis (e.g., *threshold*) and the slope of the function (how rapidly performance changes with changes in stimulus values). Often, only a location measure is required, but there are some investigations whose goal is a more complete description of the underlying function, requiring both location and slope measures. The procedures themselves involve two separate parts: placement of trials along a stimulus axis and analysis of the data obtained to extract characteristics of the underlying psychometric functions.

The challenge of adaptive psychophysics is to make relevant observations on the psychometric function with maximum efficiency without sacrificing accuracy. Adaptive methods of measurement have been developed with the goal of preserving accuracy and reliability, while maximizing efficiency and minimizing subject and experimenter time. This article will trace the development of modern adaptive procedures, with special attention to the strengths and weaknesses of three of the most commonly used methods. In addition, some special applications will be discussed, including the use of these procedures to monitor changes in performance that are due to learning or attentional lapses by subjects, as well as some characteristics of adaptive methods that become important with application to more complex applications or in studying multidimensional stimuli, such as speech. The reader is directed to Treutwein (1995) for an excellent, quantitative description of psychometric functions and adaptive techniques.

ORIGINS OF ADAPTIVE PSYCHOPHYSICAL PROCEDURES

Although adaptive procedures have been in use in some form for many years (see Levitt, 1992, for a brief discus-

sion), the systematic application of adaptive algorithms to the measurement of sensory function may be traced to clinical roots. The development of the new discipline of audiology to test hearing during and after World War II included a method devised by Hughson and Westlake (1944) for searching quickly for an auditory threshold by starting with an inaudible sound level and increasing the level until a positive response was exhibited by the patient. Hughson and Westlake emphasized the importance of sound levels beginning below a listener's threshold, then increasing until the threshold of sound was reached. Their procedure underwent minor modifications by Carhart and Jerger (1959), who proposed that an auditory test should begin at a relatively high level in order to demonstrate the sound for the listener. The level would then drop in fairly large steps until the sound became inaudible (signaled by a negative response from the subject), when the level would increase in a search for threshold. Once the listener indicated that the sound was audible (a positive response), the level would drop again, and another increasing threshold search would be initiated. The final threshold was taken to be the average level, during an increasing series, at which the listener indicated that the sound was heard. Although bearing a superficial similarity to one of the classical psychometric methodologies called the *method of limits*, this clinical method differs in that the threshold search always occurs on the ascending trials and the descending levels are used only as *subject preparation* and to get into position for the ascending threshold search. In the classical method of limits, both ascending and descending trials are used to identify where a response changes from one of audibility to inaudibility or vice versa, and the threshold is estimated from those boundary levels.

Another early adaptive procedure that has enjoyed widespread use both clinically and in research applications is Bekesy tracking, originally developed by the auditory scientist whose work in cochlear modeling won him the Nobel prize in 1962. In this procedure, a mechanical arm with a pencil attached is driven by a patient listening to tones changing continuously in intensity and, sometimes, in frequency. As the patient indicates that the tone is heard, the pen draws lower on an intensity-scaled graph, and the level of the tone decreases; if the patient indicates that the tone is not heard, tone level increases, accompanied by marks higher on the graph. The threshold, defined as that intensity at which a tone is just barely heard, is taken to be the midpoint of the up-and-down tracings of the mechanical pen.

The more commonly used adaptive procedures in research today emerged from these clinical beginnings and from methodological research in the 1940s to the 1960s. Dixon and Mood (1948) were among the first to systematically investigate the characteristics and statistical properties of simple adaptive procedures that search for threshold through a combination of increasing and decreasing stimulus steps, responding to negative and positive subject responses. This type of procedure has come to be called a *staircase* and forms the basis for a great deal of psychometric testing used today. Staircase procedures differ from earlier clinical techniques in that they collect a number of threshold estimates from both ascending and descending series of trials. They refine the method of limits by not requiring responses to a complete set of levels and by responding with changes in direction of the staircase after a change in the subject's response.

MODERN ADAPTIVE METHODS

The common characteristics of currently used adaptive methods are the collection of subject responses to each trial, with a systematic manipulation of the stimulus level along the experimental dimension of interest. Each method results in a series of stimulus levels presented over the course of the experiment, along with the associated subject responses. Experimental variables that may impact the results of the methodology include the amount of difference between stimulus values presented (the step size), the initial starting value of the stimulus, the process that guides the sequence of presentation levels on each trial (the tracking algorithm), and the decision for ending the process (the stopping rule). The general goal of each procedure is to measure characteristics of the subject's performance over the shortest amount of time, without sacrificing accuracy. Each method may be most appropriate in a given experimental situation, and there is a substantial literature comparing the abilities of these methods to provide bias-free results with high reliability.

Adaptive methodologies that currently enjoy widespread use may be placed into three general categories, defined by their systems for placing trials along a stimulus array and by the manner in which each estimates a final outcome. The first category to be described is called *parameter estimation by sequential testing* (PEST), and it is characterized by an algorithm for threshold searching that changes both step sizes and direction (i.e., increasing and decreasing level) across a set of trials. A second type of adaptive procedures has been called *maximum-likelihood procedures* but their more general characteristic is that sets of stimulus–response trials are fit with some type of ogival function and subsequent trial placement and threshold estimation is taken from those fitted functions. Finally, a common form of adaptive procedures, known generically as *staircase* procedures, will be described. For each of these three categories of procedures, an example of a tracking history will be shown, in order to understand the differences in the manner in which the adaptive rules lead to an estimate of threshold.

Parameter Estimation by Sequential Testing

The PEST procedure, first described in 1967 by Taylor and Creelman, uses changes in step size to focus the adaptive track ever more finely, stopping the track when the estimate has been adequately defined. The final estimate is simply the final value determined by the trial placement procedure. The PEST algorithm is designed to place trials at the most efficient locations along the stimulus axis in

order to increase measurement precision while minimizing the number of trials required to estimate a threshold.

Figure 2 shows a typical PEST adaptive threshold track, modified from Hall (1981), carried out according to the suggestions of Taylor and Creelman (1967). Note that an initial level and a step size are selected to begin a track. After each presentation at a fixed level, a statistical test is applied to indicate whether performance at that level is better or poorer than the targeted performance level (e.g., 75% correct detections). Once that determination is made, the level may change by the current step size, and a series of presentations occurs at the new level, again testing after each presentation whether the level should be changed. In Figure 2, an initial step size of 8 dB is used; after four presentations, the level changes; nine presentations at the new level are needed to determine that the current level is too low, and the track moves back up, but this time with a step size half as large. The next change in direction (occurring at Trial 21) produces another halving of step size. Further changes in step size (always according to PEST rules) occur throughout the track, which terminates when a criterion step size is reached. The level specified by this final step size is taken as the final threshold value. The important characteristics of this type of threshold track are that the step sizes change according to rule throughout the track, so that the track excursions tend to become smaller as a threshold value is approached, and that the final threshold estimate is taken to be the final value in the track, without specifically considering performance on previous trials.

The rules for implementation of PEST were originally outlined by Taylor and Creelman (1967), but many subsequent authors have proposed modifications. The original rules include when to change levels, a process for deciding the next level involving a step size changing throughout the track according to rule, a stopping rule based on the approach of the decreasing step size to a criterion value, and a rule for estimating the final threshold measure, typically the last level indicated by the tracking rules. Taylor and Creelman also described a metric, termed the *sweat factor,* that could be used to evaluate the efficiency of a given psychometric procedure. Likening the sweat factor to a measure of the amount of "work" necessary to reach a certain level of precision in the measuring algorithm, they defined the sweat factor as the product of the number of trials and the variance of the measures. Through simulations, they determined the variance of the PEST thresholds and the mean number of trials necessary to achieve that level of variability. Comparing results of those realistic simulations with an ideal sweat factor (generated from a simulated threshold device with complete statistical knowledge of the probabilities associated with each stimulus level) produced a measure of efficiency of the PEST procedure, which Taylor and Creelman calculated to be about 40%–50%.

PEST was designed to be as efficient as possible in the placement of trials along an array of stimulus levels and to force convergence of an adaptive track on a given performance level (i.e., threshold) as rapidly as is consistent with accuracy and reliability of measurement. The original PEST procedure called for trial placement throughout the track based on a statistical determination of performance at the current level, in comparison with expected performance at a targeted level, and a threshold estimate that was simply the final value of the track. Although some modifications to the original PEST tracking rules were suggested by Findlay (1978) in order to encourage more rapid convergence of the track, later developments emerging from the use of PEST have changed its two basic characteristics of trial placement and threshold estimation.

Hall (1981) proposed a *hybrid* procedure that followed PEST rules for trial placement along the stimulus axis, but instead of taking the final value of the track as threshold, at the end of the track, he used performance on all trials to construct a psychometric function, from which a threshold value was extracted. The value of this hybrid procedure was that efficient trial placement could proceed in a prescribed

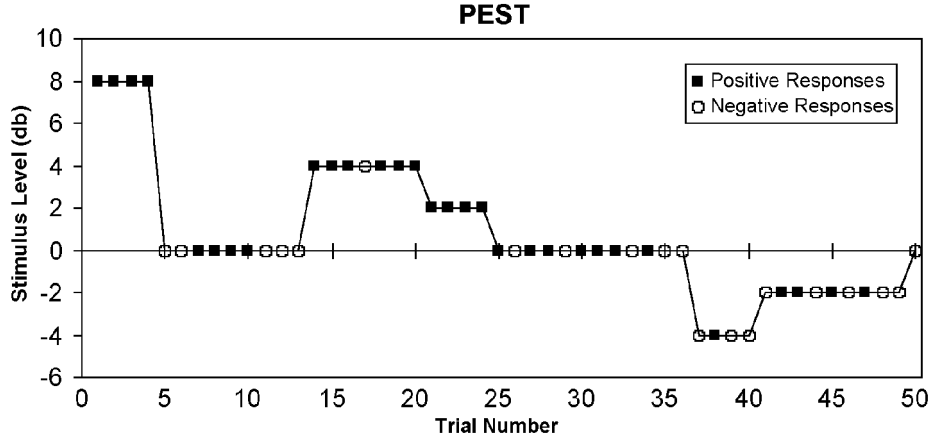

Figure 2. Adaptive track following the PEST procedure. These decibel values are relative to an arbitrary threshold of 0 dB, shown with the horizontal line. These data are modified from Figure 1 in Hall (1981).

manner according to the PEST rules, but, in the end, all the data gathered during the procedure were used to construct the final psychometric function. Through simulations and experimental trials with human subjects, Hall (1981) demonstrated that the hybrid procedure could overcome many of the shortcomings identified in previous use of PEST: The procedure was relatively tolerant of subject lapses, not affected significantly by inappropriate starting levels or step sizes, and provided estimates of both a threshold value and a slope of an assumed psychometric function.

Maximum-Likelihood Adaptive Procedures

Although Hall's (1981) hybrid procedure changed the PEST method of final threshold estimate, further modifications of PEST changed the rules for stimulus placement as well. These modifications may be classified into a second category of adaptive procedures, characterized by stimulus placement on each trial, driven by consulting the current best estimate of the entire underlying psychometric function after every stimulus–response trial. As the adaptive track grows in length, the estimated function becomes better defined by the collection of data points generated from previous trials. After each trial, the set of stimulus levels and the proportion of correct responses associated with each level are combined to form a psychometric function, as is shown schematically in Figure 1. The individual points are fitted with an ogival function of some kind (Figure 1 shows a logistic function) and a current estimated threshold level is extracted. A new psychometric function is generated after each trial or set of trials, and subsequent trials are placed at a targeted performance level on the most up-to-date function. A maximum-likelihood fitting algorithm is typically used with this type of procedure.

The link between the original PEST and the maximum-likelihood adaptive procedures may be clearly seen in papers by Pentland (1980) and Watson and Pelli (1983). Pentland developed what he called "the best PEST," to take advantage of the strength of the maximum-likelihood procedures in the context of a PEST adaptive track. Assuming a logistic psychometric function with a given slope, Pentland's procedure seeks to maximize the information provided by each trial by placing levels at the most current estimate of the 50% point on the assumed psychometric function. In simulation comparisons with the original PEST and two other adaptive procedures, Pentland's maximum-likelihood procedure proved to be the most efficient, requiring the least number of trials to reach a given level of precision. In Pentland's best PEST, levels are set according to a fitted function after each trial (using all previous trials), and a fixed number of trials is presented. The threshold estimate is simply the last value estimated as the 50% point of the ultimate psychometric function. Watson and Pelli, in their QUEST procedure, advocated the use of all information available from previous trials in the track, supplemented by prior knowledge (from the literature, previous experiments, etc.) to set the next test level. However, a distinction was made between the use of prior information to drive the track and the final estimate of a threshold, which used only the data within the track. For an adaptive track consisting of 128 trials, Watson and Pelli report an efficiency of 84% for their QUEST procedure, as compared with 40%–50% efficiency for the original PEST.

Maximum-likelihood adaptive procedures are attractive to investigators because they make full use of all trials in an experiment in order to determine a threshold, rather than estimating threshold only from the levels visited at the end of an adaptive track, as in the original PEST procedure. In most applications of these procedures, both a function shape (e.g., a logistic or Weibull function) and a slope value must be assumed, so that from trial to trial, the function moves its location along a stimulus level array in order to find the function leading to a threshold estimate. Most of the development of these procedures has occurred in the context of vision or auditory research, but Linschoten, Harvey, Eller, and Jafek (2001), in this issue, have demonstrated the value of maximum-likelihood adaptive methods in studying taste and smell. They assumed a logistic psychometric function and reported that the methods worked well in estimating thresholds with precision and speed. Although for well-studied psychophysical tasks, information concerning the function underlying performance may be known from the literature, additional nonadaptive measures might be necessary to establish the function before maximum-likelihood adaptive procedures may safely be implemented. This was the approach taken by Saberi and Green (1997) to evaluate the use of maximum-likelihood adaptive procedures in some measures of auditory discrimination of time and frequency.

An illustration of this type of adaptive methodology may be taken from Green (1993), who developed a maximum-likelihood adaptive procedure involving a yes–no psychometric task that promises highly efficient trial placement and threshold estimation. Green's method is similar to the QUEST procedure of Watson and Pelli (1983), as well as to some other implementations, but Green's procedure does not carry as many prior assumptions as does QUEST, and has a less theoretically based scheme for the placement of trials. In Green's procedure, a particular psychometric function is assumed (e.g., a logistic function), and a range of stimulus values thought to include the threshold to be estimated is identified, perhaps through pilot work or consulting the literature. A set of candidate psychometric functions is computed on the basis of the assumed form of the function and the possible stimulus values. Each of the candidate functions is fitted to all the data collected to that point after each stimulus presentation, and the likelihood associated with each function is computed. The most likely psychometric function is then visited at the target performance level to determine the stimulus level to be used on the next trial, followed by another updating of the candidate function probabilities. The final estimate of threshold is extracted from the most likely psychometric function after some number of trials or in accord with some stopping rule. Figure 3 shows a typical adaptive

track, following Green's (1993) maximum-likelihood procedure. The track is characterized by a wide excursion over the first few trials, but a rapid convergence to a threshold stimulus level. Green asserted that a reliable threshold estimate could be generated using this method in as few as 12 trials; Leek, Dubno, He, and Ahlstrom (2000), describing a stopping rule based on a criterion variability in the adaptive track, reported that, typically, 24 trials would produce highly reliable threshold estimates. In the example depicted in Figure 3, the threshold level appears to stabilize after about 20–25 trials. The procedure, developed as a yes–no task, has been extended to a forced-choice procedure by Dai and Green (1992) in a study of auditory intensity discrimination, and for frequency and intensity discrimination by He, Dubno, and Mills (1998). Further assessment of the use of this procedure for different experimental tasks and with different types of subject populations has also been reported by Leek et al. (2000).

Staircase Procedures

Both PEST and the maximum-likelihood procedures involve fairly complex stimulus placement rules and, in some cases, development of threshold estimates from the tracking data. The maximum-likelihood procedures also require the assumption of a particular form of the underlying psychometric function, which is not well established for some psychometric tasks. The simplicity and flexibility of adaptive staircases have led to their adoption as the procedures of choice in many laboratories. These methods generally use the previous one or more responses within an adaptive track to select the next trial placement, then provide a threshold estimate in a variety of ways, most commonly by averaging the levels at the direction reversals in the adaptive track (i.e., the turnaround points). Simple up–down staircases call for a reduction in stimulus level when the subject's response is positive (e.g., "I hear a tone") and an increase in stimulus level when the response is negative. Figure 4A shows an example of a simple up–down adaptive track. Beginning at a level above threshold, positive responses lead to continued decreases in stimulus level until a negative response occurs. This triggers a reversal in the direction of the track, and levels on subsequent trials increase until the next change in response. The simple up–down staircase procedure targets the 50% performance level on a psychometric function that extends from 0% correct performance at chance to 100% correct performance. In other words, the track targets the stimulus level for which the probability of a correct response equals the probability of an incorrect response or, equivalently, the level at which the track would move up or down on the stimulus axis with equal probability. The value of this type of procedure is in the very few assumptions necessary for its implementation. In contrast to the maximum-likelihood methods, no form of the psychometric function need be assumed, and there is no need for complicated computation and fitting procedures between trials. Furthermore, in contrast to PEST, the trial placement, step size, and stopping decisions are all relatively simple and straightforward. The only necessary assumption for use of these methods is a monotonic relationship between stimulus levels and performance levels.

Levitt (1971) described a general transformation procedure for targeting specific locations on a psychometric function. In the transformed methods, instead of a track level change in response to every trial, as mandated for the simple up–down procedure targeting the midpoint of the psychometric function, sequences of positive or negative responses are determined that result in an equal probability of the track's moving in either direction. For the simple up–down procedure, both the positive and the negative sequences consist of one trial, and the track level moves after each response, targeting the 50% performance level. To target a higher performance level, the sequence for a downward movement may be two or more positive responses, and the sequence for an upward movement may remain at one negative response. This example is the extensively used *two-down, one-up* procedure, which targets the 70.7% level on the psychometric function. Recalling that the probability of the down sequence must equal the probability of an up sequence, we see that a positive response to two consecutive trials must occur in order to move the track downward. If p is the probability of a pos-

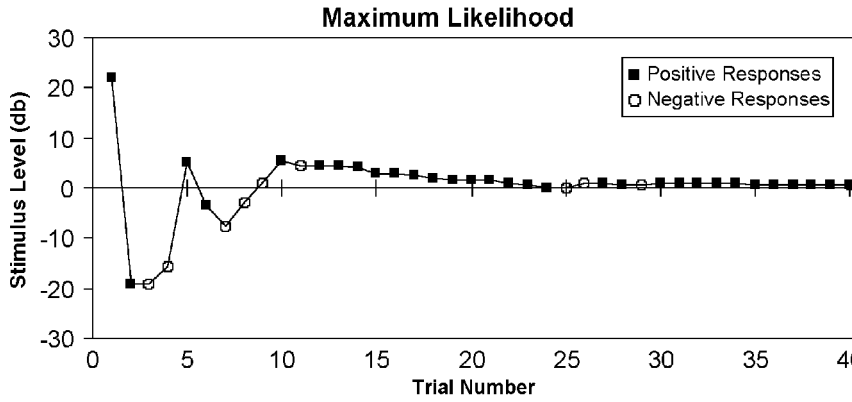

Figure 3. Adaptive track following a maximum-likelihood adaptive procedure, as developed by Green (1993). Decibel values are relative to the arbitrary threshold of 0 dB.

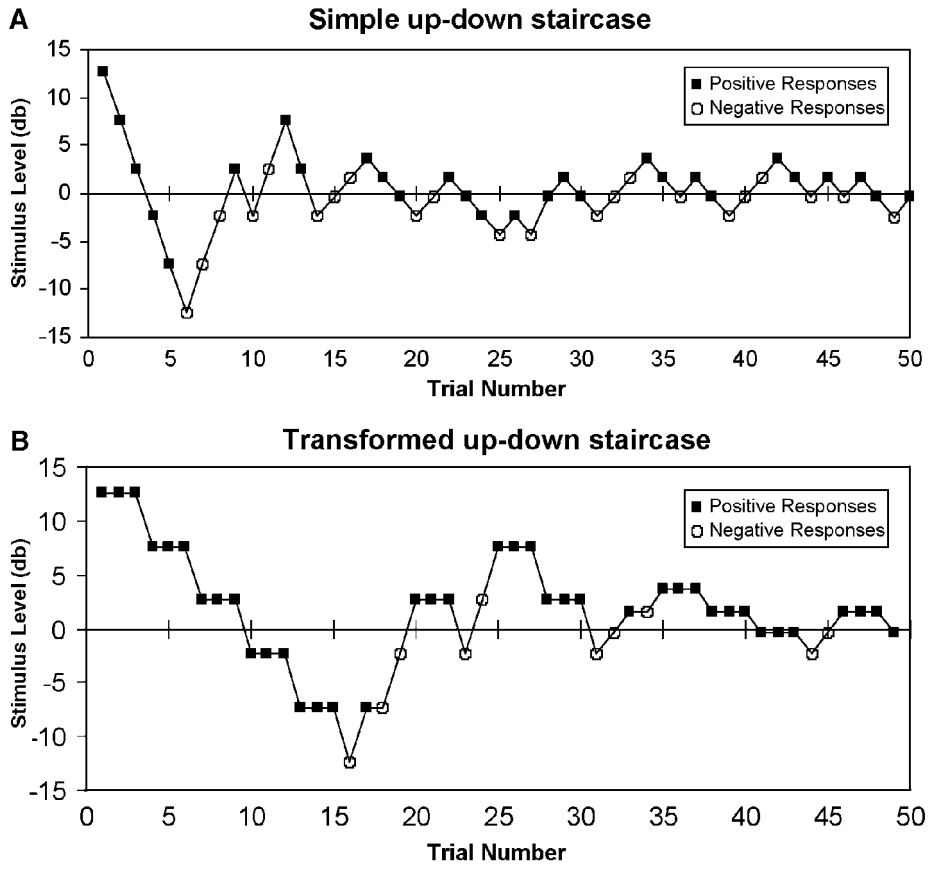

Figure 4. Adaptive tracks following a staircase procedure. (A) Simple up–down staircase; (B) transformed up–down staircase, following a three-down, one-up algorithm.

itive response on a given trial, then $p \times p$ must equal .50, and therefore the target probability is $\sqrt{.5} = .707$. Similarly, a three-down, one-up transformation leads to a performance target of .794 (i.e., $p^3 = .50$; the cube root of.50 is.794), as is shown in the example track in Figure 4B. As in the simple up–down staircase, the threshold search starts above threshold, but in this case, a decrease in stimulus level requires three sequential positive responses. A reversal in the track occurs after one negative response, and again, three positive responses are required to begin another descending run of trials. In his 1971 article, Levitt outlined a number of possible transformations, along with their target performance levels. One obvious implication is, of course, that the more complicated the sequence rule, the more trials typically required in an adaptive track in order to reach an estimate of threshold.

Although the transformed up–down methods are widely used, one restriction that has been noted is that only a small number of target levels can be estimated. Kaernbach (1991) described a simple up–down procedure that could be used to estimate performance at many more target levels than allowed by the transformed methods, by varying the step sizes used in the two different staircase directions. The value of Kaernbach's procedure, as he described it, was in the simplicity of the algorithm, relative to the sometimes quite elaborate rules necessary for the transformed procedures, and its ability to target any performance level, not just those that could be estimated with a specific sequence of up and down trials. In Kaernbach's simple up–down weighting procedure, a performance level is analyzed according to the desired ratio of up to down steps, and the stimulus level is changed after every trial. Kaernbach described an example of targeting 75% correct performance with a ratio of up to down step sizes of $(1-p)/p$ or, in this case, .25/.75, or 1/3. In order to target that point on the psychometric function, the stimulus level should be changed upward after an incorrect response and downward after a correct response, and the size of the upward step should be three times the size of the downward step. Using Monte Carlo simulations, Kaernbach (1991) reported a modest savings of about 10% of experimenter time with the weighted procedure over a more "traditional" staircase, using a two-down, one-up rule (i.e., targeting 70.7% correct). Rammsayer (1992) evaluated Kaernbach's (1991) weighted method, using human subjects in an auditory temporal discrimination task. He reported that, at the beginning of adaptive tracks, the weighted up–down method was more efficient than a

transformed fixed-step-size procedure but that, for tracks longer than about 40 trials, there was little difference in efficiency between the two procedures. Because there was a rather large difference in step size between the up and the down steps, Rammsayer noted that his subjects reported an awareness of the direction of the track, which might bias the outcome of the procedure. Rammsayer suggested that this problem could be reduced by interleaving more than one threshold track. However, it might be pointed out that any savings in experimental time would be lost if more than one track were found to be necessary for reasons other than efficiency. Kaernbach's (2001a) article in this issue uses this weighting procedure in an evaluation of threshold tracking, using an *unforced-choice* method. He argues that, although the psychometric properties of including "don't know" in the array of subject responses indicate only a small improvement over the forced-choice selections, subjects are generally more comfortable not having to indicate an answer when they are very unsure.

ESTIMATES OF SLOPE FROM ADAPTIVE PROCEDURES

Adaptive methods have tended to focus on measuring one point on the psychometric function in order to estimate a threshold or location of the function along a stimulus axis. However, in many cases, the slope of the function may be useful in fully defining the shape of the function or, for theoretical or clinical reasons, in establishing the relationship between rates of change in performance level and stimulus level. In his seminal paper, Levitt (1971) discussed the optimal placement of levels along the stimulus axis. If a measure of the mean of a psychometric function is desired as an estimate of threshold, it is most efficient to place trials as near as possible to the midpoint of the function. However, the staircase procedures may be used to estimate a slope of a psychometric function by placing trials on each side of the mean of the function. There have been a number of investigations of adaptive procedures meant to evaluate how precisely and accurately the slope, perhaps in addition to a threshold, can be determined from an adaptive procedure.

Three slightly different procedures have been proposed recently that are designed to efficiently and precisely identify the slope and threshold of the psychometric function underlying subject performance. These procedures bear some similarity to the earlier described maximum-likelihood procedures, in that, starting with a set of candidate psychometric functions (either explicitly or implicitly described), trial placement occurs at the most likely threshold value, a response is collected, and then that response is included with all other responses collected in the experimental run to generate a new set of candidate functions. The next presentation level is located at the predicted performance level most likely (at that point in the experiment) to reflect threshold (or some other targeted parameter). In the three procedures described below, meant to focus on a slope estimate, the candidate functions are not explicitly defined, but rather a two-dimensional probability distribution is updated after each response, and the two dimensions represent the two parameters defining the psychometric function.

Watt and Andrews (1981) described a procedure that was intended to maintain the advantages of a method of constant stimuli in developing a good estimation of the underlying psychometric function, using a probit fit to the data, but employing an adaptive change in stimulus presentation levels in order to increase efficiency over the method of constant stimuli. Probit (Finney, 1971) is a method of fitting a cumulative normal function to a set of psychometric data, using a maximum-likelihood criterion. Each data point to be included in the fit is weighted according to its binomial variability, and the number of trials placed at that stimulus level. Thus, points on the psychometric function that are calculated from many trials are assumed to be more reliable and, therefore, are more important in the fitting calculations. In Watt and Andrews's procedure, a few stimulus levels are selected for testing from a larger set of predetermined values, a number of trials are presented using only those levels, and then a function is fit using the probit method. A new set of stimulus levels is selected on the basis of the probit-fitted psychometric function, and a further block of trials is presented. The threshold and spread (slope) of the cumulative-normal psychometric function assumed by the probit fit converge on the values that best reflect the subject's performance on the sets of trials. Watt and Andrews advocated the use of this procedure to improve the efficiency of measuring a complete psychometric function, such as one could generate with the method of constant stimuli, without the need to assume a slope value.

More recently, King-Smith and Rose (1997) reported a method specifically designed to measure the slope of the psychometric function. They agree with Levitt (1971) that the placement of trials on the psychometric function can be selected in order to maximize efficiency of measurement either of slope or of threshold, but not of both. If threshold is the target, for maximum efficiency, trials should be placed near the target performance level on the psychometric function, and the closer to the correct performance level, the more efficient will be the measurement. However, to maximize efficiency in measuring slope, points that better define the spread of the psychometric function are appropriate. If the function itself is a symmetric function about its midpoint, two points equidistant from the midpoint of the function should be selected. An unbalanced function will require a slightly different placement of trials. King-Smith and Rose developed an adaptive method for maximum efficiency in measuring either slope or threshold, making use of the binomial variability associated with each probability level on an assumed psychometric function. Stimuli for each trial are placed with the goal of maximizing efficiency by minimizing the variability of estimates after each set of trials. The method is adaptive in that a stimulus level is determined from a probability density function generated from previous trials. The level most likely to correspond to threshold (or a

slope value) is presented. The response at this level is used to update the probability of a given response as a function of the true threshold, expressed as a likelihood function. The information provided by the likelihood function is combined with the initial probability function by Bayesian multiplication, resulting in an updated function describing the probability that each intensity is the threshold after the response to that stimulus. This new probability function is then used to begin the next trial and to estimate a threshold for optimum placement of the subsequent trial. This process can be applied to simultaneously converge on a best threshold and best slope by using two-dimensional probability density functions and likelihood functions to generate the next stimulus level. King-Smith and Rose reported that these methods result in relatively high efficiency for the measurement of both threshold and slope, which can be improved to the extent that prior knowledge of either of these parameters may be incorporated or that assumptions concerning one of them may allow experimental focus on the other.

King-Smith and Rose (1997) noted that it is somewhat more difficult to get precision in the slope parameter than in a threshold parameter. Kontsevich and Tyler (1999) developed an adaptive process similar to that of King-Smith and Rose, reporting that their procedure could produce reasonable precision in the estimate of threshold in about 30 trials for a 2AFC task but that about 300 trials were required in order to estimate slope with similar precision. As with the earlier study, the key to estimating a threshold and slope is to use each trial to update the posterior probability of a given two-dimensional probability distribution. Kontsevich and Tyler identified a potential problem with selecting the minimum variance of this distribution in that the two dimensions of threshold and slope are incommensurate and, therefore, some weighting convention must be imposed. Instead, these authors proposed a different *cost* factor, which could be minimized in order to determine the next stimulus level for presentation. This minimized factor corresponds to maximum information gain after each trial. Using computer simulations and psychophysical experiments with humans, this procedure was shown to converge to threshold and slope values within a relatively small number of trials.

Leek, Hanna, and Marshall (1992) investigated the utility of using performance on all the trials that make up an adaptive threshold track to generate a full psychometric function from which a threshold and slope value could be extracted, reminiscent of Hall's (1981) earlier suggestion of a hybrid PEST–maximum-likelihood procedure. Leek et al. (1992) used computer simulations to determine the precision of estimate of slopes that could be accomplished from a staircase procedure designed to track threshold at a particular performance level. The simulations were meant to determine whether both slope and threshold could be obtained by simply reconstructing a psychometric function on the basis of the trial-by-trial performance within an up–down transformed staircase track. A procedure that is optimized to produce threshold measurements might also provide slope information, with little loss in precision. Experimental runs were generated following selected staircase algorithms by consulting a known psychometric function on each trial to determine a response. At the end of an adaptive track, the trial-by-trial data were summarized according to performance at each level visited by the track and then fit to a psychometric function of the same form as the one consulted in the simulation. Thresholds generated by the tracking algorithm were compared with thresholds extracted from the original and reconstructed psychometric functions. Slope estimates from the fitted psychometric functions were compared with those underlying performance on the adaptive trials. The functions reconstructed from the trial-by-trial data were accurate reflections of the underlying functions as long as the tracks were at least 200 trials long. Shorter tracks resulted in estimates of psychometric function slope that were biased high (i.e., slopes too steep). It was noted that a maximum-likelihood fit of the data to the psychometric function provided greater stability of estimates than did an earlier set of analyses carried out using the probit fitting procedure (Finney, 1971). The authors cautioned against using the probit procedure when the psychometric function was transformed so that the range of performance was less than 0%–100%. The statistical properties of such a fit, described fully by McKee, Klein, and Teller (1985), alter the variability associated with each point in the transformed function, which is critical as a weighting component in the probit procedure. Thus, when the psychometric function is truncated, as in forced-choice procedures, McKee et al. recommended that the probit fit should be avoided. The Leek et al. (1992) reconstructions of psychometric functions from tracking data resulted in a finding regarding slope estimates similar to that frequently reported: It is possible to obtain accurate and reliable slope estimates from adaptive procedures, but the cost is more trials and subsequently more experiment time.

Slopes estimated from adaptive tracking procedures reported by Leek et al. (1992) tended to be too high unless the tracks were fairly long. This tendency for an overestimation of slope from adaptive methods has been observed before and has usually been attributed to an asymmetry in the distribution of trial placements along the underlying psychometric function. Typically, more trials are placed higher than the midpoint of the function than are placed lower. However, Kaernbach (2001b), in this issue, argues that the source of the slope overestimation is not the poor distribution of trials, but, instead, may be found in the adaptive algorithms themselves. Kaernbach shows that, for trials placed identically to trial placement within an adaptive track but presented in random order instead of according to the sequence of the algorithm, slopes are not overestimated, lending support to his notion that the sequential aspects of the trial placement, according to adaptive rules, is the source of the slope bias.

Determination of slope requires some assumptions about the form of the psychometric function. There are a number of ogival functions to choose from, and criteria for an experimenter's choice may depend on theoretical or computational issues. Strasburger (2001a), in this issue, points out that comparisons across studies and tasks are hampered by the variety of psychometric functions that are selected and provides formulas for converting among a number of the most commonly used functions. It is suggested in that paper that for comparison of results across studies, the maximum slope of a function, or the slope at the point of inflection of the ogival function, would be a useful metric. In a second paper in this issue (Strasburger, 2001b), character recognition is measured using a 10-alternative forced-choice procedure with a maximum-likelihood/PEST procedure, and the maximum-slope metric is used to compare results across studies.

Also in this issue may be found three articles addressing the best way to sample and fit the psychometric function. Although probit fitting is commonly used, it is not always the most appropriate choice, because of changes in binomial variability with truncation of the function for forced-choice procedures and because it assumes a particular form of the function—that is, the cumulative normal. Two papers by Wichmann and Hill (2001a, 2001b) take up the issues of how best to sample the function and how to determine the goodness of fit of the assumed function. Miller and Ulrich (2001) describe a method for fitting the psychometric function that makes no assumption about the underlying distribution, as does the more commonly used probit analysis.

VIOLATION OF ASSUMPTIONS IN ADAPTIVE METHODS

One of the strongest arguments for using adaptive procedures for the rapid and accurate estimates of characteristics of psychometric performance is that there are very few restrictions that must be accommodated. Two such commonly accepted requirements, however, involve stability of the measurement over time and the monotonic relationship between stimulus strength and performance. Most experimenters acknowledge that absolute stability of the underlying psychometric function generally is not a realistic assumption, since subjects typically experience some perceptual learning during the course of an experiment, reducing the true threshold, or occasionally have lapses in attention that may serve to increase the measured threshold. Changes in threshold across the measurement track may also result in shallower calculated slopes of the underlying function. Violations of a second assumption, monotonicity of the psychometric function, may occur if members of the stimulus set under examination are not homogenous along a given stimulus dimension. Stimuli with greater complexity and dimensionality—such as speech, for example—are likely not homogenous and, therefore, pose special problems when tested adaptively. There have been attempts to monitor violations of the assumptions of stability of the functions and homogeneity of the stimuli and to assess the costs of such violations.

Tracking Threshold Changes With Multiple Adaptive Tracks

Although a potential problem in the implementation of adaptive procedures, it is also one of its benefits, as identified by Levitt (1971) and earlier papers, that adaptive tracking procedures may be used to follow changes in the psychometric function occurring during the course of an experiment. For example, as perceptual learning occurs over many trials, the threshold may be seen to gradually (or suddenly) improve, as reflected in the shapes of the threshold tracks. Leek and Watson (1984) used this method to trace the improvements in detection of individual tones embedded within a 10-tone pattern. Ten interleaved threshold tracks, one testing each of the 10 pattern components, were examined to determine how the tone frequency and temporal placement within the pattern affected the improvements in detection thresholds. In contrast to improvements in threshold over time, adaptive tracks may also signal subject fatigue or distraction over the course of an experiment. Hall (1983) suggested a method of identifying a shift in a subject's threshold, by comparing the subject's response on each trial with the response predicted from the estimated psychometric function underlying performance. To the extent that the difference between obtained and predicted performance is close to zero for each trial in the track, the estimated psychometric function is taken to be stable throughout the track. Leek, Hanna, and Marshall (1991) also proposed a method for determining whether the true threshold of a subject was shifting during the time of its measurement, thereby producing an inaccurate threshold estimate. Their approach was to investigate the psychometric properties of an unstable underlying psychometric function by simulating systematic changes in function location and comparing the variability within and across two interleaved adaptive tracks. The logic was that if a psychometric function were changing over time, the variability in levels visited within a single track would exceed the variability observed between tracks on trials occurring close together in time. In addition to using these two sources of variability to monitor whether the function is shifting in time, Leek et al. (1991) showed how the across-track variability may be used to generate an estimate of the slope of the underlying function. In computer simulations and human subject data, both the slope estimates and the stability-monitoring procedure were shown to work well as long as the shift in thresholds did not occur so rapidly that the tracks could not follow the changes.

Nonmonotonic Psychometric Functions and Heterogenous Stimuli

Early in the course of development of modern adaptive methods, Levitt and Rabiner (1967) attempted to apply an adaptive procedure to the measurement of levels of speech necessary for a given level of performance. Typically, speech testing is accomplished by presenting a list of

speech stimuli at a given level and asking listeners to repeat the stimuli they heard. The speech commonly consists of standardized lists of monosyllables, consonant–vowel nonsense syllables, sentences, or running discourse. Often, it is desired to measure performance at a number of different presentation intensities or in different types and levels of competing noise. Speech recognition as a function of increasing level or increasing signal-to-noise ratio is an ogival psychometric function. Speech testing typically focuses on some portion of the rising part of the function. Levitt and Rabiner applied an up–down staircase and blocks of stimuli at each tested level to determine 50% correct identification of the speech. Bode and Carhart (1973) extended these findings by developing what they called the *doublet* procedure, using a transformed up–down adaptive method. They ran two sequential tracks, targeting the 29.3% and the 70.7% correct identification signal-to-noise level. The average of the final threshold levels from each track constituted an estimate of 50% correct performance. Steele, Binnie, and Cooper (1978) used the doublet adaptive procedure to study the impact of visual cues (lip reading) on tests of speech understanding, using monosyllabic words as stimuli.

Although adaptive procedures continue to be widely used in measuring speech recognition, the nature of speech stimuli creates greater variability in measurement than is observed in testing more homogenous sets of stimuli, such as tone detection or discrimination. When stimuli are homogenous in all characteristics except the one under test (e.g., the level of a pure tone in a fixed noise), the procedure can work well. Similarly, if a set of speech stimuli is homogenous in all factors affecting intelligibility, the level of the stimuli either in quiet or in noise may be used in an adaptive procedure. However, if the stimuli are not homogenous, the tracking procedure may be compromised. For example, a series of easily heard monosyllables may drive the level of the track low, but a subsequent presentation of an inherently more difficult word will occur at a level that is too low, and the track may not truly reflect overall performance. Heterogeneity within stimulus sets thus leads to inappropriate placement of trial levels, greater variability in the track, and possible confusion to the subject. Dirks, Morgan, and Dubno (1982) noted this difficulty when testing identification of monosyllables and of spondee words (i.e., two-syllable words with equal stress on each syllable) in a group of normal hearing and a group of hearing-impaired subjects. The speech level was held constant throughout an adaptive track, whereas the level of multitalker babble was varied according to an adaptive algorithm designed to target 29.3%, 50%, or 70.7% correct word identification. In these procedures, a simple up–down procedure was first initiated to find the correct range of performance levels for a given individual, after which either the simple up–down algorithm was continued (50% correct) or one of the transformed up–down algorithms was implemented to target the other two performance levels. Dirks et al. established thresholds for each target and stimulus set but reported that performance of the tracks was indeed more variable when the stimuli were monosyllables, rather than the more homogenous set of spondee words. This increased variability was especially striking for the group of hearing-impaired listeners, who likely experienced other sources of variability associated with their hearing loss as well. It is important, therefore, in using adaptive methods with sets of speech stimuli, to control the heterogeneity of the stimuli to the extent possible and to be alert to violation of the monotonicity assumption across trials of an adaptive track.

COMPARISONS OF ADAPTIVE PROCEDURES

Are any of these adaptive procedures better than others, and how do the procedures interact with other psychometric experimental choices? There have been a number of papers comparing the accuracy and efficiency of the procedures described above in measuring thresholds and slopes of psychometric functions. These comparisons have been made by using computer simulations of experimental tests and, in some cases, evaluating the performance of human listeners.

Shelton, Picardi, and Green (1982) evaluated an adaptive staircase, a maximum-likelihood procedure, and PEST in collecting data from human subjects. In each case, they chose parameters for the adaptive procedures that were commonly used in practice, measuring human performance on a simultaneous- and a forward-masking auditory task. Although the procedures did not produce large differences, there were some characteristics of each that might suggest one choice or another under certain circumstances. For short adaptive runs (i.e., less than 30 trials), both the staircase and the maximum-likelihood procedures resulted in slightly biased threshold estimates, although the bias could be mostly overcome by randomizing starting levels for each adaptive run. The maximum-likelihood procedure, however, was the only one of the three methods that could converge within about 10 trials, even with a slightly biased threshold. Shelton et al. suggested that this procedure might be most useful in testing infants and animals, where the thresholds must be gathered rapidly. However, they pointed out that the maximum-likelihood method may be particularly difficult for inexperienced listeners, because there are very few suprathreshold trials afforded to the subjects.

Kollmeier, Gilkey, and Sieben (1988) used a mathematical model, as well as human data, to compare two staircase procedures and the PEST procedure, with both a 2AFC and a three-alternative forced-choice (3AFC) experimental task. They also evaluated both model and human listeners on a set of fixed level (nonadaptive) procedures. Their human listeners were all experienced in the task, which was detection of a signal embedded in noise—that is, simultaneous masking. The model predicted similar thresholds from adaptive and nonadaptive procedures, but in practice, human listeners actually produced slightly bet-

ter thresholds in adaptive procedures. The model predicted that a 2AFC procedure used in a staircase targeting 71% correct would be the least efficient procedure, whereas the 79% 3AFC staircase would be the most efficient. The human data also supported the latter as the most efficient procedure, but results were somewhat variable. One of the problems identified with human data, in contrast to the modeled data, was that the model underestimated the variability produced by human listeners. The authors suggested that the reason for this might be an underlying psychometric function that is slowly varying and suggested that variability in the tracks may be partitioned into a rapid trial-to-trial variability, combined with variability contributed from a slowly varying function. Hall (1983) had earlier made such a suggestion, and it later was taken up in simulations of unstable psychometric functions by Leek et al. (1991). Kollmeier et al. suggested that the variability might be controlled by combining thresholds from short tracks, rather than using long, perhaps varying, tracks to estimate thresholds. Hicks and Buus (2000) agreed with this notion, finding more consistent thresholds from interleaving several short tracks than from following one long track involving the same total number of trials. In summary, Kollmeier et al. reported that, if no other experimental considerations dictate otherwise, the most efficient combination of methods is the 79% 3AFC. This conclusion has been reached by a number of other authors.

Kollmeier et al. (1988) also found that thresholds from adaptive tracks in human performance tended to be biased low (i.e., better thresholds) than would be expected by fixed trials. Even though some simulation studies suggest that thresholds should be similar from fixed-level and adaptive procedures, there are consistent reports of thresholds of human subjects being better in adaptive procedures meant to target the same level as fixed-level methods. Taylor, Forbes, and Creelman (1983) reported this in describing their comparisons of PEST procedures with fixed-level procedures. Stillman (1989) compared thresholds measured with the adaptive procedures and thresholds from fixed-level tests, finding that the adaptive methods always produced thresholds that were lower, just as was reported by Kollmeier et al. (1988) and Shelton et al. (1982). In other words, the adaptive procedures always overestimated subjects' performance, producing better (lower) thresholds than were evidenced in a nonadaptive task.

In Stillman's (1989) study, both inexperienced and experienced subjects were used, and comparisons were made for results from adaptive and nonadaptive procedures and for two staircase procedures and a PEST adaptive procedure. The task was a 2AFC detection of a 1-kHz tone in a bandpass noise centered at 1 kHz. Results indicated similar thresholds for the 79% staircase and the PEST procedures (targeting 80% correct) and similar variability within the two staircase procedures. Shelton and Scarrow (1984) also measured performance of human listeners to determine whether some experimental choices were better than others. All their listeners were inexperienced, and they used separate groups of 10 listeners each for each of four conditions, staircase and maximum likelihood, using both a 2AFC and a 3AFC procedure in each. The task was detection of a tone in noise. Together with Shelton et al. (1982), these authors reported that thresholds were essentially equivalent for all the 2AFC procedures tested (staircase, PEST, and maximum-likelihood) and for the 3AFC staircase and maximum-likelihood procedures. They did observe some differences in variability and efficiency across the procedures, noting that the 3AFC staircase provided the most stable thresholds across adaptive runs but that the maximum-likelihood method produced stability early in a run. Therefore, according to these authors, if practice trials are not possible or the number of trials is limited, the maximum-likelihood procedure should be favored.

Schlauch and Rose (1990) primarily used simulations with a small set of human data to investigate the use of staircase procedures with 2-, 3-, and 4AFC tasks. They measured both efficiency (variability) and threshold bias as a function of number of intervals, step size, and the target of the adaptive track (equivalently, the decision rule for changing stimulus levels). They identified less variability in threshold measurements as the number of intervals increased—especially, from 2 to 3 intervals, less so between 3 and 4 intervals—and for the higher performance target (79% vs. 71%). Even taking into account the greater experimental time necessary to present the larger number of intervals, the 3AFC and 4AFC procedures were still more efficient than the 2AFC. They also reported greater variability in threshold estimates for larger step sizes. The 2AFC 71% target was more biased (i.e., identifying better performance) than the 4AFC 79%, and there was more bias for larger step sizes, especially for the 71% 2AFC procedure. Schlauch and Rose suggested that this bias was a result of behavior near chance performance and the effects of guessing. They also found no improvements in performance of the methods for adaptive tracks longer than 100 trials. By fitting the trial-by-trial data, using a probit method (Finney, 1971), the thresholds recovered some of the bias that was associated with all the adaptive procedures. In order to improve efficiency and reduce bias, these authors recommended fitting the trial-by-trial data in the adaptive track to estimate a threshold and the use of small step sizes in the tracking procedure. Although the 4AFC procedure gave the best efficiency and the least bias, the time taken to present four intervals on each trial may strain the memory of subjects and may, in the end, increase experimental time even though fewer trials might be necessary.

In summary, there is little to recommend any of the three reviewed psychometric procedures from the standpoint of the performance of the methods themselves. It seems clear that some experimenter selections of various implementations of the methods may increase or decrease the bias and reliability of the procedures. In particular, the 2AFC task is generally a poor choice, particularly when paired with a staircase target of 71%. McKee et al. (1985) provided a clear description of the impact of truncating the 0%–100% psychometric function when using a forced-choice procedure such as the 2AFC. Instead of the func-

tion's spanning a large range from chance performance at 0% correct to perfect performance at 100%, these truncated functions result from increased chance levels (e.g., 50% for 2AFC, 33.3% for 3AFC), and therefore, the range of the psychometric function is decreased. The variability associated with each point of the function, however, contributes to the bias and variability of measurement according to the binomial distribution. Therefore, in general, points falling lower than the midpoint of the truncated functions generally have greater variability. McKee et al. suggested that measurements are likely to be more reliable if they are on the upper side of the midpoint of the function. The binomial variability of the truncated psychometric functions may account for the relatively poorer psychometric performance of the 2AFC 71% combination (target lower than the 75% midpoint of the 2AFC function), with better performance when the combination of forced-choice task and target performance level lead to trials placed higher on the function. Green (1990) addressed a similar point, arguing that the best placement of stimulus trials was near the top end of the psychometric function.

SUMMARY AND CONCLUSIONS

Three categories of adaptive procedures were reviewed. PEST procedures do not require assumptions about the shape of the underlying psychometric function and provide a rapid and systematic convergence on a threshold. Maximum-likelihood procedures for placing trials at optimal stimulus levels and for providing threshold and slope estimates are computationally intensive and require assumptions regarding the shape of the underlying function. However, they converge on targeted values very quickly and make good use of all the data collected in a track. Staircase methods require very few assumptions and have fairly simple algorithms for placement of stimuli and estimation of threshold values. They may support an estimate of slope, so long as sufficient trials are presented.

Some of these methods have slight advantages over others, given particular experimental circumstances. For example, when testing must be accomplished very quickly, as in testing animals or infants, the faster converging maximum-likelihood procedures might offer some benefit over longer staircase procedures. There is strong consensus, however, that the popular 2AFC procedures do not have desirable statistical properties, particularly when paired with adaptive procedures that target relatively low performance levels (i.e., below the midpoint of the psychometric function) and should be avoided. Finally, stimuli tested in adaptive procedures should have the characteristic of homogeneity and a monotonic relationship between stimulus level and performance level. This has been shown to be problematic (although not fatal) in testing some kinds of speech recognition adaptively.

Adaptive methods offer high precision and reliability in psychometric testing, at a significant savings in time over nonadaptive testing. Over the last 50 years, refinements and evaluations of these procedures have shown the way to a selection of experimental variables and parameters that result in little cost for the savings in time. Although there are inherent biases in some of the methods, these can be mostly compensated by a thoughtful consideration of experimental techniques and parameters.

REFERENCES

Bode, D. L., & Carhart, R. (1973). Measurement of articulation functions using adaptive test procedures. *IEEE Transactions on Audio and Electroacoustics*, **AU-21**, 196-201.

Carhart, R., & Jerger, J. F. (1959). Preferred method for clinical determination of pure-tone thresholds. *Journal of Speech & Hearing Disorders*, **24**, 330-345.

Dai, H., & Green, D. M. (1992). Auditory intensity perception: Successive versus simultaneous across-channel discriminations. *Journal of the Acoustical Society of America*, **91**, 2845-2854.

Dirks, D. D., Morgan, D. E., & Dubno, J. R. (1982). A procedure for quantifying the effects of noise on speech recognition. *Journal of Speech & Hearing Disorders*, **47**, 114-123.

Dixon, W. J., & Mood, A. M. (1948). A method for obtaining and analyzing sensitivity data. *Journal of the American Statistics Association*, **43**, 109-126.

Fechner, G. T. (1860). *Elemente der Psychophysik* (Vol. 1). Leipzig: Breitkopf & Härtel. [Also available as *Elements of Psychophysics*. New York: Holt, Reinhart & Winston, 1966.]

Findlay, J. M. (1978). Estimates on probability functions: A more virulent PEST. *Perception & Psychophysics*, **23**, 181-185.

Finney, D. J. (1971). *Probit analysis* (3rd ed.). Cambridge: Cambridge University Press.

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.

Hall, J. L. (1983). A procedure for detecting variability of psychophysical thresholds. *Journal of the Acoustical Society of America*, **73**, 663-667.

He, N.-J., Dubno, J. R., & Mills, J. H. (1998). Frequency and intensity discrimination measured in a maximum-likelihood procedure from young and aged normal-hearing subjects. *Journal of the Acoustical Society of America*, **103**, 553-565.

Hicks, M. L., & Buus, S. (2000). Efficient across-frequency integration: Evidence from psychometric functions. *Journal of the Acoustical Society of America*, **107**, 3333-3342.

Hughson, W., & Westlake, H. (1944). Manual for program outline for rehabilitation of aural casualties both military and civilian. *Transactions of the American Academy of Ophthalmology & Otolaryngology*, **48** (Suppl.), 1-15.

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

Kaernbach, C. (2001a). Adaptive threshold estimation with unforced-choice tasks. *Perception & Psychophysics*, **63**, 1377-1388.

Kaernbach, C. (2001b). Slope bias of psychometric functions derived from adaptive data. *Perception & Psychophysics*, **63**, 1389-1398.

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.

Kollmeier, B., Gilkey, R. H., & Sieben, U. K. (1988). Adaptive staircase techniques in psychoacoustics: A comparison of human data and a mathematical model. *Journal of the Acoustical Society of America*, **83**, 1852-1861.

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

Leek, M. R., Dubno, J. R., He, N.-J., & Ahlstrom, J. B. (2000). Experience with a yes–no single-interval maximum-likelihood procedure. *Journal of the Acoustical Society of America*, **107**, 2674-2684.

Leek, M. R., Hanna, T. E., & Marshall, L. (1991). An interleaved tracking procedure to monitor unstable psychometric functions. *Journal of the Acoustical Society of America*, **90**, 1385-1397.

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

Leek, M. R., & Watson, C. S. (1984). Learning to detect auditory pattern components. *Journal of the Acoustical Society of America*, **76**, 1037-1044.

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

Levitt, H. (1992). Adaptive procedures for hearing aid prescription and other audiologic applications. *Journal of the American Academy of Audiology*, **3**, 119-131.

Levitt, H., & Rabiner, R. L. (1967). Use of a sequential strategy in intelligibility testing. *Journal of the Acoustical Society of America*, **42**, 609-612.

Linschoten, M. R., Harvey, L. O., Jr., Eller, P. M., & Jafek, B. W. (2001). Fast and accurate measurement of taste and smell thresholds using a maximum-likelihood adaptive staircase procedure. *Perception & Psychophysics*, **63**, 1330-1347.

McKee, S. P., Klein, S. A., & Teller, D. Y. (1985). Statistical properties of forced-choice psychometric functions: Implications of probit analysis. *Perception & Psychophysics*, **37**, 286-298.

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

Pentland, A. (1980). Maximum-likelihood estimation: The best PEST. *Perception & Psychophysics*, **28**, 377-379.

Rammsayer, T. H. (1992). An experimental comparison of the weighted up–down method and the transformed up–down method. *Bulletin of the Psychonomic Society*, **30**, 425-427.

Saberi, K., & Green, D. M. (1997). Evaluation of maximum-likelihood estimators in nonintensive auditory psychophysics. *Perception & Psychophysics*, **59**, 867-876.

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.

Steele, J. A., Binnie, C. A., & Cooper, W. A. (1978). Combining auditory and visual stimuli in the adaptive testing of speech discrimination. *Journal of Speech & Hearing Disorders*, **43**, 115-122.

Stillman, J. A. (1989). A comparison of three adaptive psychophysical procedures using inexperienced listeners. *Perception & Psychophysics*, **46**, 345-350.

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.

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.

Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research*, **35**, 2503-2522.

Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, **33**, 113-120.

Watt, R. J., & Andrews, D. P. (1981). APE: Adaptive probit estimation of psychometric functions. *Current Psychological Reviews*, **1**, 205-214.

Wichmann, F. A., & Hill, N. J. (2001a). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, **63**, 1293-1313.

Wichmann, F. A., & Hill, N. J. (2001b). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception & Psychophysics*, **63**, 1314-1329.

(Manuscript received June 22, 2001;
revision accepted for publication September 25, 2001.)