Pergamon

0042-6989(95)00016-X

Vision Res. Vol. 35, No. 17, pp. 2503–2522, 1995

0042-6989/95 $29.00 + 0.00

Minireview

Adaptive Psychophysical Procedures

BERNHARD TREUTWEIN *

Received 8 July 1993; in revised form 12 January 1995

Improvements in measuring thresholds, or points on a psychometric function, have advanced the field of psychophysics in the last 30 years. The arrival of laboratory computers allowed the introduction of adaptive procedures, where the presentation of the next stimulus depends on previous responses of the subject. Unfortunately, these procedures present themselves in a bewildering variety, though some of them differ only slightly. Even someone familiar with several methods cannot easily name the differences, or decide which method would be best suited for a particular application. This review tries to illuminate the historical background of adaptive procedures, explain their differences and similarities, and provide criteria for choosing among the various techniques.

Psychometric functions Psychophysical threshold Binary responses Sequential estimate Efficiency Yes–no methods Forced-choice methods

INTRODUCTION

The term *psychophysics* was invented by Gustav Theodor Fechner, a 19th-century German physicist, philosopher and mystic. For him psychophysics was a mathematical approach to relating the internal psychic and the external physical world on the basis of experimental data. Fechner (1860) thereby developed a theory of the measurement of internal scales and worked out practical methods, the now *classical* psychophysical methods, for estimating the *difference threshold*, or *just noticeable difference (jnd)*, the minimal difference between two stimuli that leads to a change in experience. Today, the threshold is considered to be the stimulus difference that can be discriminated in some fixed percentage of the presentations, e.g. 75%. Fechner's original methods were as follows:

The method of constant stimuli: a number of suitably located points in the physical stimulus domain are chosen. These stimuli are repeatedly presented to the subject together with a comparison or standard stimulus. The cumulative responses (different or same) are used to estimate points on the psychometric function, i.e. the function describing the probability that subjects judge the stimulus as exceeding the standard stimulus.

The method of limits: the experimenter varies the value of the stimulus in small ascending or descending steps starting and reversing the sequence at the upper and lower limit of a predefined interval. At each step the subject reports whether the stimulus appears smaller than, equal to or larger than the standard.

The method of adjustment is quite similar to the method of limits and is only applicable when the stimulus can be varied quasi-continuously. The subject adjusts the value of the stimulus and sets it to apparent equality with the standard. Repeated application of this procedure yields an empirical distribution of the stimulus values with apparent equality which is used to calculate the point of subjective equivalence (PSE).

In general, each of these three methods suffers from one or more of the following deficits:

- absence of control over the subject's decision criterion;
- the estimates may be substantially biased;
- no theoretical justification for important aspects of the procedure;
- a large amount of data is wasted since the stimulus is often presented far from threshold where little information is gained.

In the last 35 yr, different remedies for each of these deficits have been suggested. The first two drawbacks and the lack of theory were addressed by the application of detection and choice theory to psychophysics (Luce, 1959; 1963; Green & Swets, 1966; Macmillan & Creelman, 1991). Efficiency of data acquisition was improved by using computers in psychophysical laboratories: psychophysical stimuli are generated online, and the tests are administered, scored, and interpreted by computer in a single session. Thereby the stimulus presentations are concentrated around the presumed location of the threshold.

This review gives a survey of the different methods for accelerated testing which have been proposed during recent decades. The arrangements used are sophisticated modifications of the method of constant stimuli and the method of limits. Apart from speeding up threshold mea-

*Institut für Medizinische Psychologie, Ludwig-Maximilians-Universität München, Goethestr. 31, D-80336 München, Germany [*Email* bernhard@tango.imp.med.uni-muenchen.de].

surement, some of the methods try to address the lack of theoretical foundation, while others remain purely heuristic arrangements.

BASIC CONCEPTS

Experimental designs

Experiments based on Fechner's classical methods measure *discrimination*, the ability to tell two stimuli apart. A special case of discrimination experiments is often called *detection*: if one of the two stimuli is the null stimulus (like average luminance in a contrast sensitivity experiment) the discrimination experiment can be called a detection paradigm. In both cases we deal with classical *yes–no* designs, where the subject has to decide whether the stimuli of the two classes are the same (*no* response) or different (*yes* response). These classical designs are in contrast to *forced choice* designs, where the subject has to identify the spatial or temporal location of a target stimulus. There is no restriction for adaptive procedures to be used in yes–no or forced choice designs, but the problems considered in this article will be restricted in two other aspects:

i The response domain is limited to experiments which have binary outcomes.

ii The stimulus domain has to be represented by a one-dimensional continuum. This does not restrict the problem to continuous variables but leaves out the following two classes of problems. First, problems where the stimulus domain is a nominal scale; e.g. classification of polyhedra or similarity of words. Second, it excludes more-dimensional problems where two or more parameters are varied conjointly, e.g. in the context of colour discrimination (MacAdam, 1942; Silberstein & MacAdam, 1945) or in joint frequency/orientation discrimination (Treutwein, Rentschler & Caelli 1989). In two-dimensional problems the single value of a threshold corresponds to a closed curve around a reference stimulus delineating a non-discrimination area. An adaptive procedure would have to track this curve and determine the geometrical parameters of the curve from the subject's responses.

Psychometric function

Plotting the cumulative responses of an experiment with binary outcomes against the stimulus level results in the psychometric function. Throughout this article *percentage yes* responses (yes–no design) and *percentage correct* responses (forced choice design) will be used synonymously in the context of psychometric functions.

An example of a psychometric function with results from a forced-choice experiment with nine spatial alternatives is given in Fig. 1. Here, the percentage correct assignments of the stimulus location has been plotted against the stimulus level, which in this case was the duration of a temporal break in one of nine simultaneously displayed stimuli. The plotted results are cumulative data of 35 sessions, i.e. repetitions of the experiment with the same stimulus setup. Thresholds for break duration, i.e. double-pulse resolution, was measured by use of YAAP, an adaptive procedure of the Bayesian type (see below; for experimental details see Treutwein & Rentschler, 1992).

A problem with almost every real observer can be seen in Fig. 1: even at stimulus levels far higher than the threshold, which was in this design 55.$\bar{5}$% correct (at stimulus level 24), real subjects exhibit a tendency not to notice the stimulus, i.e. to have *lapses* — some people use the term *rate of false negative errors*. Similar behaviour also occurs below the threshold when subjects sometimes give a *yes* response. The probability of such responses is termed the *guessing rate* or the *rate of false positive errors*. In a yes–no design this behaviour probably reflects noise in the sensory system whereas in forced choice designs *correct* responses below threshold are normal: the subjects are forced to give a localization answer, even when they did not perceive anything; in this case the best they can do is to guess. In forced choice experiments with an *unbiased*[†] observer these responses occur with a probability of $\frac{1}{n}$ if the subject has to choose from n alternatives. The lapsing rate p_l and the guessing rate p_g can sometimes be estimated from the collected data in a subsequent analysis of the responses, but usually both have to be prespecified by the experimenter. In Fig. 1 the guessing rate of 9.4% was estimated by the percentage of correct responses at stimulus level 1 and the lapsing rate of 2.4% was estimated by the percentage of correct responses after collapsing the results from all presentations at levels between 37 and 99. This collapsed percentage correct value and the corresponding number of trials are marked in Fig. 1 as ◆.

Usually the guessing rate p_g is accounted for by applying Abbott's formula which yields an adjusted rate of correct answers $\psi^*(x)$ from the actually measured rate $\psi(x)$:

$$\psi^*(x) = \frac{\psi(x) - p_g}{1 - p_g}. \tag{1}$$

Sometimes this formula is extended to include the lapsing rate p_l:

$$\psi^*(x) = \frac{\psi(x) - p_g}{1 - p_g - p_l}. \tag{2}$$

Solving for $\psi(x)$ yields the following:

$$\psi(x) = p_g + (1 - p_g)\,\psi^*(x)$$

or

$$\psi(x) = p_g + (1 - p_g - p_l)\,\psi^*(x).$$

It is important to keep in mind that the responses at any fixed stimulus level are binomially distributed. This implies that the variability of the percentage correct measures, and therefore the precision whith which percentage correct can be measured, depends on both, the number of

[†]An unbiased observer is a hypothetical subject who distributes his or her guesses equally between the different alternatives. This is not necessarily the case for a real observer.

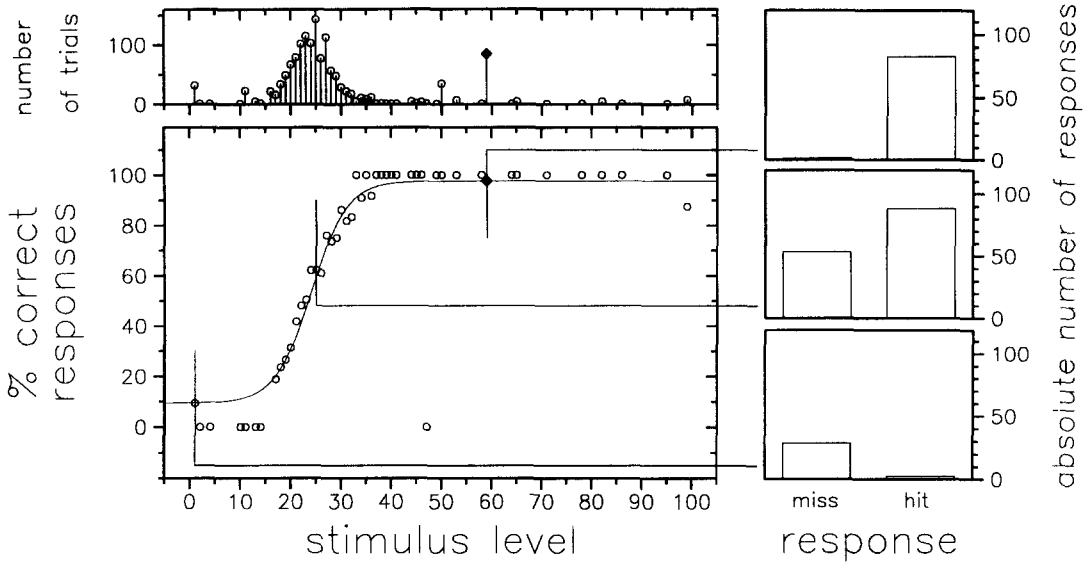

FIGURE 1. **Binomial Responses and the Psychometric Function.** Illustration of the psychometric function and the underlying binomial distribution at fixed stimulus values. The left part of the figure shows: top, a histogram of the number of presentations; bottom, the percentage of correct responses with a nonlinear regression line of a logistic psychometric function. The three subplots on the right-hand side depict the actual number of correct/incorrect answers at three specific stimulus values thereby illustrating the binomial distribution of these responses.

trials *and* the unknown "true" percentage correct at that stimulus level. The variance of a success probability p_s, when the underlying responses are binomially distributed, is $\mathrm{Var}(p_s) = p_s(1 - p_s)/n$, where p_s = % correct/100, and n is the number of trials at that stimulus level. Therefore, when fits of theoretical models to the data are sought, a measure of the variability of the unadjusted rate of correct answers should be used as weighting factor for the adjusted rate even when an adjusted psychometric function [equation (1) or (2)] is used.

Due to the presence of guessing and lapsing behaviour, the psychometric function $\psi(x)$ is *not* a cumulative probability distribution though it looks very similar, i.e. in almost every real experiment a psychometric function does not fulfil the asymptotic requirements for a cumulative probability distribution $F(x)$,

$$\lim_{x\to-\infty} F(x) = 0 \text{ and } \lim_{x\to+\infty} F(x) = 1$$

but instead fulfils (3)

$$\lim_{x\to-\infty} \psi(x) = p_g \text{ and } \lim_{x\to+\infty} \psi(x) = p_l,$$

i.e. the psychometric function has the guessing and lapsing rates p_g and p_l as asymptotic values.

Threshold

The goal of threshold experiments is to find a stimulus difference that leads to a preselected percentage of correct responses, i.e. to a preselected level of the subject's performance, i.e. the threshold. A probability value ϕ is set and the corresponding stimulus level x_ϕ is sought. This corresponding stimulus value $x_\phi = \theta$ is called the *threshold.** For yes–no designs the threshold is usually chosen to be the 50% point, the point where *same* and *different* responses are equally likely. This type of yes–no threshold is called the *point of subjective equivalence*. For forced choice designs the threshold is often chosen to halve the interval between the guessing and lapsing rate, i.e. $\phi = (p_l - p_g)/2$.

Adaptive procedures

As pointed out by Falmagne (1986) the difference between the classical and the adaptive methods is that in the former, the stimulus values which will be presented to the subject are completely fixed before the experiment; whereas in the latter, they depend critically on the responses of the subject: the stimulus presented on trial n depends on one, several or all of the preceding trials. Put in a more formal way, the value of the stimulus level presented in an adaptive psychophysical experiment at trial n is considered as a stationary stochastic process, i.e. the stimulus value x_n which is presented on trial n depends on the outcome of the preceding trials. Since the subjects' responses form a stochastic process, the stimulus values also constitute one. Therefore the stimulus level at trial n will be denoted by the random variable $\mathbf{X}_n$ and the subject's response by the random variable $\mathbf{Z}_n$. The actual values of the response $\mathbf{Z}_n$ are coded as $z_i = 0$ for a miss (*same* response in a yes–no design or *incorrect* assignment in a forced choice design), and $z_i = 1$ for a hit (*different* or *correct* response). By definition of the psychometric function, we have

*This kind of threshold is called an *empirical threshold* and it is unrelated to those of the threshold theories; an empirical threshold can be measured either in terms of detection theory, e.g. d', or in terms of threshold theory, e.g. percentage correct (see Macmillan & Creelman, 1991, Chap. 4 and 8).

$$\text{Prob}\{\mathbf{Z}_n = 1|\mathbf{X}_n\} = \psi(\mathbf{X}_n)$$
$$\text{and} \qquad (4)$$
$$\text{Prob}\{\mathbf{Z}_n = 0|\mathbf{X}_n\} = 1 - \psi(\mathbf{X}_n).$$

This means that at any fixed stimulus level the responses of the subject are binomially distributed (see also the right hand insets in Fig. 1).

With these formal definitions, an adaptive procedure is given by a function $\mathcal{A}$ which combines the presented stimulus values $\mathbf{X}_n$ and the corresponding responses $\mathbf{Z}_n$ at trial n and preceding trials with the target probability ϕ to yield an optimal stimulus value $\mathbf{X}_{n+1}$, which is then presented on the next trial

$$\mathbf{X}_{n+1} = \mathcal{A}(\phi, n, \mathbf{X}_n, \mathbf{Z}_n, \ldots, \mathbf{X}_1, \mathbf{Z}_1). \qquad (5)$$

The implication of the stationarity of the stochastic process for a psychophysical experiment is that consecutive presentations should be statistically independent (e.g. by interleaving different runs for independent parameters in one experimental session; also see the Discussion).

Performance of a method

Psychophysical procedures should be evaluated in terms of cost and benefits. The currency in which psychometric procedures are bought is the patients' or subjects', and the experimenter's time, i.e. the number of trials required to achieve a certain accuracy. An empirical threshold is a statistic, an estimate of a theoretical parameter. In other words, the threshold is a function of the data, which is a summary measure that depends on the results of a set of trials. The relevance of this statistic is assessed by:

i Bias, or systematic error, i.e. is the *estimated* threshold on average equal to the *true* threshold?
ii Precision, i.e. some measure inversely related to the variability, or the random error. If the threshold is measured repeatedly, how much variation is to be expected?
iii Efficiency, i.e. how many trials are required to achieve a certain precision?

Before considering precision, bias and efficiency in more detail, I would like to make two remarks about the minimum number of trials necessary to obtain accurate estimates:

i The more parameters are to be estimated, the more trials are necessary.
ii The more the target probability ϕ deviates from $\frac{1}{2}$, where the binomial distribution of the subject's responses has the highest variance, the more trials are necessary.

Bias. The difficulty of evaluating the bias of a particular psychophysical method is that in any real experiment one does not know the value of the true threshold, i.e. in real experiments, the experimenter can never determine how large the bias is. Evaluating the bias of a method therefore can be done only in simulations. King-Smith, Grisby, Vingrys, Benes and Supowit (1994) have pointed to the difference between the *measurement* bias and the *interpretation* bias. Measurement bias is the difference between the true value and the average estimated value. Interpretation bias is the result of an inverse question: given a single estimated value of a threshold, one may ask what values of real thresholds could have given rise to this threshold estimate. More specifically, what are the relative probabilities of different real thresholds which could have given rise to this threshold estimate, and what is the weighted average of these real thresholds? I will come back to the question of interpretation bias in the section on Bayesian methods and the interpretation of the *a-posteriori* distribution.

If the value of the *true* threshold θ_{true} is known, the measurement bias $b_{\hat{\theta}}$ of r estimated thresholds $\hat{\theta}_r$ can be defined in the following way:

$$b_{\hat{\theta}} = \frac{1}{r}\sum_r(\theta_{\text{true}} - \hat{\theta}_r) = \theta_{\text{true}} - \mu_{\hat{\theta}}, \qquad (6)$$

where r is the number of estimates considered in this case, i.e. the number of repeated sessions or runs, and $\mu_{\hat{\theta}}$ is the mean of these best estimates.

Precision. The precision $\kappa_{\hat{\theta}}$ of r estimated thresholds $\hat{\theta}_r$ can be defined (see Taylor, 1971) as the inverse of the variance of the best threshold estimates $\hat{\theta}$ of a particular method, i.e.

$$\kappa_\theta = \frac{1}{\sigma^2_{\hat{\theta}}} = \frac{r-1}{\sum_r(\hat{\theta}_r - \mu_{\hat{\theta}})^2}, \qquad (7)$$

where $\mu_{\hat{\theta}}, \sigma^2_{\hat{\theta}}$ are the mean and the variance of the best estimates and r is the same as in equation (6).

Efficiency. Taylor and Creelman (1967) and Taylor (1971) have defined the *sweat factor K* as a measure of the efficiency of a psychophysical procedure. It is the product of $\sigma^2_{\hat{\theta}}$, the variance of the best threshold estimate, and n, the fixed number of trials, which were necessary to obtain that estimate

$$K = n\sigma^2_{\hat{\theta}} = n\frac{\sum_r(\mu_{\hat{\theta}} - \hat{\theta}_r)^2}{r-1}. \qquad (8)$$

The sweat factor allows for comparison between different psychophysical methods. If an absolute measure of efficiency is desired then an ideal procedure as a standard of reference has to be assumed. Taylor (1971) proposed as a measure of an ideal procedure the asymptotic variance σ^2_{RM} of the Robbins–Monro process (see section on stochastic approximation) for a given target probability ϕ and a given number of trials n

$$\sigma^2_{\text{RM}} = \frac{\phi(1-\phi)}{n\left(\frac{d\psi(x)}{dx}\Big|_\theta\right)^2}, \qquad (9)$$

where $\frac{d\psi(x)}{dx}\Big|_\theta$ is the slope of the psychometric function at the threshold. An ideal sweat factor of an optimal procedure, according to this definition of the ideal process, is therefore given by

$$K_{\text{ideal}} = n\sigma_{\text{RM}}^2 = \frac{\phi(1-\phi)}{\left(\frac{d\psi(x)}{dx}\Big|_\theta\right)^2}, \quad (10)$$

and the efficiency of a procedure under consideration (index p) could be stated as

$$\eta_p = \frac{K_{\text{ideal}}}{K_p}. \quad (11)$$

The sweat factor and this definition of efficiency is applicable in cases where adaptive procedures are evaluated with a fixed number of trials in each session. For sequential procedures, which terminate after a prespecified confidence in the estimate is reached, the obvious measure of efficiency is the number of trials which were necessary to reach that point (see Daintith & Nelson, 1989).

An important question of efficiency is the behaviour of the procedure for different starting points, i.e. how the initial uncertainty about the location of the threshold and the variability of the final estimate are related to each other.

Constituents of an adaptive procedure

Adaptive procedures differ from the classical ones mainly in that they are designed to concentrate stimulus presentations at or near the presumed value of the threshold.

The procedure of any adaptive method can be divided into several subtasks:

(1) When to change the testing level and where to place the trials on the physical stimulus scale?
(2) When to finish the session?
(3) What is the final estimate of the threshold?

Not all procedures reviewed here explicitly specify all parts, although for any adaptive procedure this should be done in detail. Table 1 summarizes all procedures which will be dealt with in this article and gives an overview as to which author specified which subtask in the suggested procedure. These subtasks are in principle independent and can be exchanged without any loss. It is, for example, a permissible combination to use the stimulus placement from stochastic approximation, the termination criterium of YAAP and the final threshold from a probit analysis, or any other reasonable mixture. Moreover there is no restriction on changing any of these rules in the midst of a procedure.

Differences between categories of *adaptive psychophysical methods* concern what the experimenter already knows about the — in principle unknown — form of the underlying psychometric function and what she/he wants to learn about it:

(1) The psychometric function is known to be strictly monotonic but its shape is unknown. The experimenter is mainly interested in the stimulus value which corresponds to the prespecified performance.
(2) The experimenter knows that the psychometric function can be described by a function with several degrees of freedom which correspond to threshold, slope, and possibly further parameters controlling the asymptotes. The experimenter wants to estimate both the psychometric function's threshold and slope.
(3) The shape of the psychometric function is completely known, i.e. the experimenter chooses a family of curves, which are shift invariant on the stimulus axis. In short the only parameter to be estimated is the threshold.

In the first case the methodology of *non-parametric* statistics is used whereas in the latter two *parametric* models are assumed.

NONPARAMETRIC METHODS

In this section I will summarize different methods in which no parametric model for the psychometric function is used. These methods try to track a specific target value, i.e. the threshold. The only requirement for the psychometric function is monotonicity. Most of these methods could probably be considered as being special cases of stochastic approximation methods (Robbins & Monro, 1951; Blum, 1954; Kesten, 1958; see below).

Truncated staircase method

The simplest extension of the method of limits is to truncate the presentation sequence after any shift in the response category, thus avoiding the presentation of stimuli far below and above the threshold. This is the *truncated method of limits* or *simple up–down method:* After each trial the physical stimulus value is changed by a fixed the step size δ. If a shift in the response category occurs (from success to failure or vice versa), the direction of steps is changed. Every sequence of presentations, where the stimulus value is stepped in one direction, is called a run, and the final estimate is obtained by averaging the reversal points. This is sometimes called a midrun estimate. A more elaborate way to calculate the final estimate was given by Dixon and Mood (1948) who proposed a maximum likelihood estimate for the threshold. * Because numerical solutions for the maximum-likelihood estimate, which will be described below in the section on maximum-likelihood and Bayesian estimation, were unfeasible at that time, Dixon and Mood gave analytical approximations for the threshold and its variability.

The stepping rule for the simple up-down method can be formalized as follows, where the general equation (5) takes the form

$$\mathbf{X}_{n+1} = \mathbf{X}_n - \delta(2\mathbf{Z}_n - 1). \quad (12)$$

Here δ is the fixed step size. The stimulus value $\mathbf{X}_n$ is increased by δ for a failure ($\mathbf{Z}_n = 0$) and is decreased by the same amount for a success ($\mathbf{Z}_n = 1$).

The experiment starts with an "educated" guess for the first presentation $\mathbf{X}_1$ and the sequence of stimuli is deter-

*Dixon and Mood estimated the sensitivity of explosives to shock when a weight is dropped from different heights on a specimen of an explosive mixture. They already noted that the same method can be applied to threshold measurement in psychophysical research.

TABLE 1. Summary of all reviewed procedures. Entries marked with ⋆ have an underlying statistical proof, those marked with † have heuristic arguments and those marked with ‡ are based on questionable arguments. A — column entry means that this part was not specified in the original article. ML stands for maximum likelihood.

procedure		rules for			
name/author	year	changing levels	placing stimuli	stopping	final estimate
truncated staircase		every trial	equal steps δ	—	see text
Dixon & Mood	1947	every trial	equal steps	—	ML (all trials)
stochastic approximation	1951	every trial	$\frac{c}{n}$ (see text)⋆	—	—
non-parametric Up-Down	1957	r.v. (see text)	r.v. (see text)⋆	—	—
accelerated stoch. appr.	1958	every trial	$\frac{c}{m_s}$ (see text)⋆	—	—
PEST (MOUSE mode)	1967	Wald test	heuristic rules†	step size	last level
UDTR	1970	rules	rules⋆	—	last tested
PEST (RAT mode)	1975	Wald test	rules‡	step size	mean level
virulent PEST	1978	sliding Wald test	PEST rules†	step size	last tested
MOBS	1988	every trial	bisection‡	—	—
weighted Up-Down	1991	every trial	two step sizes: δ_1, δ_1	—	—
APE	1981	every 10/16 trials	1.35 SD (see text)	—	probit/2D-ML
Hall's hybrid	1981	Wald test	PEST rules†	—	2D-ML
Hall	1968	every trial	current best	no. of trials	ML
QUEST	1979	every trial	curr. best (Bayes)	no. of trials	ML
BEST PEST	1980	every trial	current best	no. of trials	ML
ML-TEST	1986	every trial	curr. best (Bayes)	χ^2 test	ML
Emerson	1986	every trial	current best	no. of trials	Bayes–mean
IDEAL	1987	every trial	current best	no. of trials	Bayes
YAAP	1989	every trial	current best	Bayes prob. interval	Bayes–mean
STEP	1990	every trial	least squares‡	no. of trials	least sqares
ZEST	1991	every trial	current best	no. of trials	Bayes–mean

mined by equation 12. Dixon and Mood (1948) showed with the assumption of an underlying cumulative normal distribution, i.e. the probability for a correct answer as a function of stimulus intensity being $p_c(x) = \mathcal{N}(x; \mu, \sigma)$ (see also Textbox 2) that the optimal step size is between 0.5σ and 2.4σ of the underlying distribution. Since σ is in general unknown this helps the experimenter only when knowledge from previous or similar experiments can be used.

An important restriction of the truncated staircase method is that it converges only to the target probability $\phi = 0.5$. In a forced choice experiment or any other setup, where this target probability is unsuitable, one of the following methods should be used: *transformed up–down methods* (Levitt, 1970), *non-parametric up-and-down experimentation* (Derman, 1957), the *weighted up–down method* (Kaernbach, 1991), or *stochastic approximation* (Robbins & Monro, 1951).

Transformed up–down method

In the *up–down transformed-response* (UDTR) method, Levitt (1970) suggested that changes of the stimulus value be made to depend on the outcome of two or more preceding trials. For example, the level is increased with each incorrect response and decreased only after two successive correct responses (1-up/2-down, or 2–step rule). The upward and downward steps are of the same size. Levitt has given a table of eight rules which converge to six different target probabilities ($\phi \in \{0.159, 0.293, 0.5, 0.707, 0.794, 0.841\}$). For the 2-step rule, the convergence point is $\phi = 0.707$. These rules are derived from the probabilities which are expected on the basis of the underlying binomial distribution and have a sound theoretical foundation.

Non-parametric up–down method

In the case $\phi \geq 0.5$, Derman (1957) suggested the following procedure:

$$\mathbf{X}_{n+1} = \mathbf{X}_n - \delta(2\mathbf{Z}_n\mathbf{S}_\phi - 1), \tag{13}$$

where $\mathbf{S}_\phi$ is a binomial random variable with $p = \frac{1}{2\phi}$. This means that for a correct answer the stimulus value is decreased by δ with a probability of $\frac{1}{2\phi}$, but eventually it can also be increased with the complementary probability. For an incorrect answer the stimulus value is always increased.

The non-parametric up–down method is based on statistical theory and has an underlying proof.

Weighted up–down method

Smith (1961) gave a hint and Kaernbach (1991) clearly formulated an extension to the truncated staircase method, where different step sizes for upward and downward steps are used. The relation between these being

$$\delta_\uparrow = \delta_\downarrow \frac{1-\phi}{\phi}, \tag{14}$$

where $\delta_\uparrow$ denotes the upward and $\delta_\downarrow$ the downward step size.

Modified Binary Search

Tyrell and Owens (1988) suggested a method which is based on the bisection method commonly used for finding a value in an ordered table (see Press, Teukolsky, Vetterling and Flannery (1992) Chap. 3.4) or for finding a root* of a function (ibid., Chap. 9). Bisection here means that a stimulus interval, which brackets the root, is halved consecutively and on each step one of the two endpoints is replaced by the bracketing midpoint. The normal usage requires that the function evaluation is deterministic. Tyrell and Owens have adopted this algorithm for a probabilistic response function and have added heuristic precautions, arguing that they are necessary because the subject's threshold is non-stationary. Most of the reasoning of modified binary search (MOBS) as applied to psychometric functions, i.e. taking into account the probabilistic nature of the subjects' responses, is heuristic and lacks a theoretical foundation.†

Stochastic approximation

Robbins and Monro (1951) have shown that for any value of ϕ between 0 and 1 the sequence given by

$$\mathbf{X}_{n+1} = \mathbf{X}_n - \frac{c}{n}(\mathbf{Z}_n - \phi) \tag{15}$$

converges to $\theta = x_\phi$ with probability 1. Here, c is a suitably chosen constant. The only necessary assumption about $\psi(x)$ is that it is a strictly increasing function. Equation (15) leads to increments in the stimulus value for misses and to decrements for hits. The step size δ depends on the initial step size c, the target probability ϕ, and the number of trials n: for $\phi = 0.5$, upward or downward steps on trial n are equal ($\delta = c/(2n)$); $\phi \neq 0.5$ leads to asymetric step sizes, i.e. an increment of size $c\phi/n$ if an incorrect answer was given and a decrement of size $c(1-\phi)/n$ for a correct answer. Both increments and decrements become smaller the longer the experiment runs, since the step size δ is proportional to c/n. This sequence‡ of stimuli is known as a *Robbins–Monro process*. The method guarantees that the sequence of stimulus values converges to the threshold when only the monotonicity of the psychometric function is granted. Although the original article neither specified a stopping criterion, nor how the final estimate is obtained, these are discussed by Dupač (1984) and Sampson (1988). A reasonable stopping criterion would be a lower limit for the step size and an obvious final estimate is the last tested level.

Accelerated stochastic approximation

Kesten (1958) suggested a method called *accelerated stochastic approximation*. During the first two trials the standard stochastic approximation equation (15) is used, but afterwards the step size is changed only when a shift in response category occurs (from correct to incorrect or vice versa):

$$\mathbf{X}_{n+1} = \mathbf{X}_n - \frac{c}{2 + m_{\text{shift}}}(\mathbf{Z}_n - \phi), \quad n > 2. \tag{16}$$

Here, m_{shift} is the number of shifts in response category. Kesten proved that sequences, which change the step size only when shifts in the response category occur, also converge to x_ϕ with probability 1 but do so with fewer trials than the Robbins–Monro process. The same remarks on the stopping criterion and the final estimate apply as for the stochastic approximation.

In automated static perimetry (Bebie, Fankhauser & Spahr, 1976; Spahr, 1975) a standard method for varying the intensity of the stimuli is the *4-2 dB strategy*, which can be interpreted as the first part of an accelerated stochastic approximation sequence.

PEST *and More Virulent* PEST

PEST, an acronym for Parameter Estimation by Sequential Testing, was suggested by Taylor and Creelman (1967). On the one hand, PEST was the first procedure where methods of sequential statistics were applied to psychophysics. On the other hand, in PEST a completely heuristic set of rules for stimulus placement was employed. The methodology of sequential statistics is used to determine the minimum number of responses which — at a given stimulus value — are required to reject the null-hypothesis that the responses are binomially distributed with a mean of the target probability.

Assume that the experimenter has picked a certain stimulus level x and has presented n trials at this level. The responses at this stimulus level are binomially distributed. The null hypothesis is $p = \psi(x)$, where $\psi(x)$ is the value of the unknown psychometric function at stimulus level x. When this level is far from the target value $\theta = x_\phi$, a simplified version of a sequential probability ratio test (SPRT) (Wald, 1947; see Textbox 1 for the original SPRT) will fail after a few trials. In this case, the actual number of correct responses is inconsistent with the assumption that the last stimulus presentation was at x_ϕ

*The point x_0 where a function $f(x)$ takes the value 0. The function $f(x) = \psi(x, \theta) - \phi$ has its root at the threshold.

†The correct adaptation of the root finding algorithm to probabilistic functions is the stochastic approximation (see Sampson 1988).

‡The proofs for stochastic approximation and its accelerated version are valid for more general sequences of decrementing the step size. Every sequence $\{a_n\}$ which fulfils the following three condition works: $\lim_{n\to\infty} a_n = 0$, $\sum_1^\infty a_n = \infty$, and $\sum_1^\infty a_n^2 = A \leqslant \infty$. The simplest example for such a sequence is $a_n = c/n$.

Let $\mathbf{Z}_n$ denote the binomially distributed random variable of the responses at a certain stimulus value. For n presentations at that fixed level, m_c denotes the number of successes and $n - m_c$ is the number of failures.
A sequential test of strength (α, β) is given by the probabilities for type I errors α (rejecting a correct hypothesis) and type II errors β (accepting a wrong hypothesis). In our case we are testing the null hypothesis $H_0(\phi = \phi_0)$ against the alternate hypothesis $H_1(\phi = \phi_1)$, where ϕ_0, ϕ_1 are two different target probabilities with $0 < \phi_0 < \phi_1 < 1$.
The probability p of obtaining the sample of n responses, with $E[m_c] = \phi n$ correct ones, where ϕ denotes the unknown probability for a correct answer at the current stimulus value, is given by

$$p = \phi^{m_c}(1 - \phi)^{n-m_c}$$

For the specific probabilities ϕ_0 and ϕ_1 at trial n we get:

$$p_0 = \phi_0^{m_c}(1 - \phi_0)^{n-m_c} \quad \text{and} \quad p_1 = \phi_1^{m_c}(1 - \phi_1)^{n-m_c}$$

After each response the discrimination value d is calculated:

$$d = \log \frac{p_1}{p_0} = m_c \log \frac{\phi_1}{\phi_0} + (n - m_c) \log \frac{1 - \phi_1}{1 - \phi_0}$$

If $d \in (\log \frac{\beta}{1-\alpha}, \log \frac{1-\beta}{\alpha})$ the presentation at the current stimulus level is continued. If $\log \frac{1-\beta}{\alpha} \leq d$ then H_0 is rejected, and H_1 is accepted, which means that the stimulus value should be increased. If $d \leq \log \frac{\beta}{1-\alpha}$ then H_0 is accepted, i.e. the experimenter can be confident with error probabilities (α, β) that the last tested stimulus value was at the target probability ϕ.

Textbox 1. Wald's sequential probability ratio test.

and the stimulus level is changed according to a set of heuristic rules (see below).

Taylor and Creelman's simplified version of the SPRT differs from the original Wald test in that a heuristic deviation limit is used in the following way. The expected number of correct answers m_c, after n_x presentations at stimulus level x for a target probability ϕ, is given by the mean of the binomial distribution $\mathcal{B}(n_x, \phi)$ with n_x repetitions and probability ϕ:

$$E[m_c] = \phi n_x .$$

The experimenter has chosen a deviation limit w such that:

$$N_b^{(\pm)} = E[m_c] \pm w = \phi n_x \pm w .$$

N_b is called the bounding number for the correct responses after n_x trials at the fixed stimulus value x. If the observed number of correct answers m_c is within this bracket, i.e.

$$m_c \in [\, N_b^{(-)}, N_b^{(+)} \,] = [\, \phi n_x - w,\ \phi n_x + w \,],$$

then testing at the current stimulus level is continued. If the actual number of correct responses m_c is outside this interval, the current stimulus level is changed accordingly. Taylor and Creelman suggested a value of $w = 1$ for a target probability of $\phi = 0.75$.

Taylor and Creelman proposed the following heuristic rules for changing the stimulus level, which have been empirically tested to track the value of the threshold:
(1) on every reversal, halve the step size;
(2) the second step in a given direction is the same size as the first;
(3) the fourth and subsequent steps in a given direction are each double their predecessor;
(4) whether a third successive step in the given direction is the same or double the second depends on the sequence of steps leading to the most recent reversal. If the step immediately preceding that reversal resulted from doubling, then the third step is not doubled, while if the step leading to the most recent reversal was not the result of a doubling, then this third step is double the second.

In Taylor and Creelman's original version, the session is terminated when the step size falls below a certain predefined value. The final estimate for the threshold is the last tested value x. This simple way to derive a final estimate is called PEST's MOUSE mode (Minimum Overshoot and Undershoot Sequential Estimation) as opposed to the RAT mode (Rapid Adaptive Tracking), which was introduced by Kaplan (1975). In PEST's RAT mode, the final estimate is derived by averaging the obtained stimulus values every 16 trials. The different modes and a slightly revised set of stepping rules can be found in more detail in Macmillan & Creelman (1991) Chap. 8

A modification of PEST by Findlay (1978) which the author claims to be faster, is called MORE VIRULENT PEST. It changes the power of the SPRT during the experimental run by letting the deviation limit w be a function of the number of presentations and the number of reversals. Findlay (1978) also suggested fitting the psychometric function to the cumulative results of all presentations.

PARAMETRIC METHODS

The methods discussed in the following sections require a prior decision about the general form of the psychometric function. This means that a special parametric template for the psychometric function, e.g. the cumulative

Normal distribution:

$$\mathcal{N}(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\, dt$$

definition range: $x \in (-\infty, +\infty)$
parameter set: $\Theta = (\mu, \sigma)$

with: $\mu \in (-\infty, +\infty)$ mean (position)
$\sigma > 0$ standard deviation
σ^2 variance.

Logistic distribution:

$$\mathcal{L}(x;\alpha,\beta) = \frac{1}{1 + \exp(\frac{\alpha - x}{\beta})}$$

definition range: $x \in (-\infty, +\infty)$
parameter set: $\Theta = (\alpha, \beta)$

with: $\alpha \in (-\infty, +\infty)$ position parameter
$\beta > 0$ spread parameter
with $\beta = \sigma/1.7$ and $\alpha = \mu$ the logistic function is a fairly good approximation to the cumulative normal. Some authors use $\beta'(\alpha - x)$ as the argument for the exponential; in this case β' is usually called the *slope* parameter.

Step function:

$$S(x;\alpha) = \begin{cases} 1 & \text{if } x \geq \alpha \\ 0 & \text{if } x < \alpha \end{cases}$$

with: α location of the step.

Note that any of the other model functions approximate a step function when a very large slope is used.

Weibull distribution:

$$\mathcal{W}(x;\alpha,\beta) = 1 - \exp\left\{-\left(\frac{x}{\alpha}\right)^{\beta}\right\}$$

definition range: $x \in (0, +\infty)$
parameter set: $\Theta = (\alpha, \beta)$
with: $0 < \beta$ form parameter
$0 < \alpha$ scale parameter

On a logarithmic x-axis α is the position and β the slope parameter (see also the Gumbel distribution below).

Alternate forms The cumulative Weibull distribution has no standard notation and is often written in other forms:

$$\mathcal{W}^{(1)}(x;\alpha,\beta,\gamma) = 1 - \exp\left\{-\left(\frac{x-\gamma}{\alpha}\right)^{\beta}\right\}$$

$$\mathcal{W}^{(2)}(x;\lambda,\beta) = 1 - \exp\left\{-\lambda x^{\beta}\right\}$$

$$\mathcal{W}^{(3)}(x;\alpha,\beta) = 1 - \exp\left\{-\frac{x^{\beta}}{\alpha}\right\}$$

Gumbel distribution:

$$\mathcal{G}(x;\alpha,\beta) = 1 - \exp\left\{-\exp\left(\frac{x-\alpha}{\beta}\right)\right\}$$

definition range: $x \in (-\infty, +\infty)$
parameter set: $\Theta = (\alpha, \beta)$
with: $\alpha \in (-\infty, +\infty)$ position parameter
$\beta > 0$ spread parameter

The Gumbel distribution can be useful in place of the Weibull when the latter is used on an intensity scale and translation invariance on a log-intensity scale is wanted. The Gumbel distribution has this property directly on the log-intensity, e.g. the dB scale.

Textbox 2. Formulas of typical psychometric functions.

normal, the Weibull, or the logistic distribution, is chosen and one or two of the free parameters of this template are estimated, namely the threshold and the slope. Examples of different psychometric function templates are shown in Textbox 2 and Fig. 2.

Estimation of threshold and slope

Hall (1981) and Watt and Andrews (1981) proposed two different methods, both of which estimate two parameters: the threshold and the slope of the psychometric function. Except for the idea of splitting the complete session into several blocks, and of estimating the psychometric function's parameters between these blocks of presentations, both use different methods for parameter estimation and stimulus placement.

Adaptive probit estimation. The approach of Watt and Andrews (1981) is based primarily on the classical method of constant stimuli but differs in that it adjusts the placement of the stimuli during the run according to the outcome of a probit analysis (Finney, 1971; Textbox 3). The session is split up into blocks of short constant-stimuli subsessions with four different stimulus values x_1, x_2, x_3, x_4. The authors originally suggested blocks of 10 presentations. The experimenter supplies educated guesses of the threshold μ_0 and spread (inverse slope) σ_0. Before the rth block the spacing of the four "constant" stimulus values is derived from the current estimates by

$$x_{\{1..4\}}^{(r)} = \{\, \mu_r - \sigma_r,\ \mu_r - \frac{c}{3}\sigma_r,\ \mu_r + \frac{c}{3}\sigma_r,\ \mu_r + \sigma_r \,\}. \quad (17)$$

According to Watt and Andrews probit analysis is of optimum efficiency when the constant $c = 1.35$ and probit is applied to data gathered with the method of constant stimuli. At the end of the second and of every subsequent block, a "rapid and slightly approximate Probit analysis" of the last two blocks is carried out to obtain new best estimates $\hat{\mu}, \hat{\sigma}$. With these best estimates a new stimulus

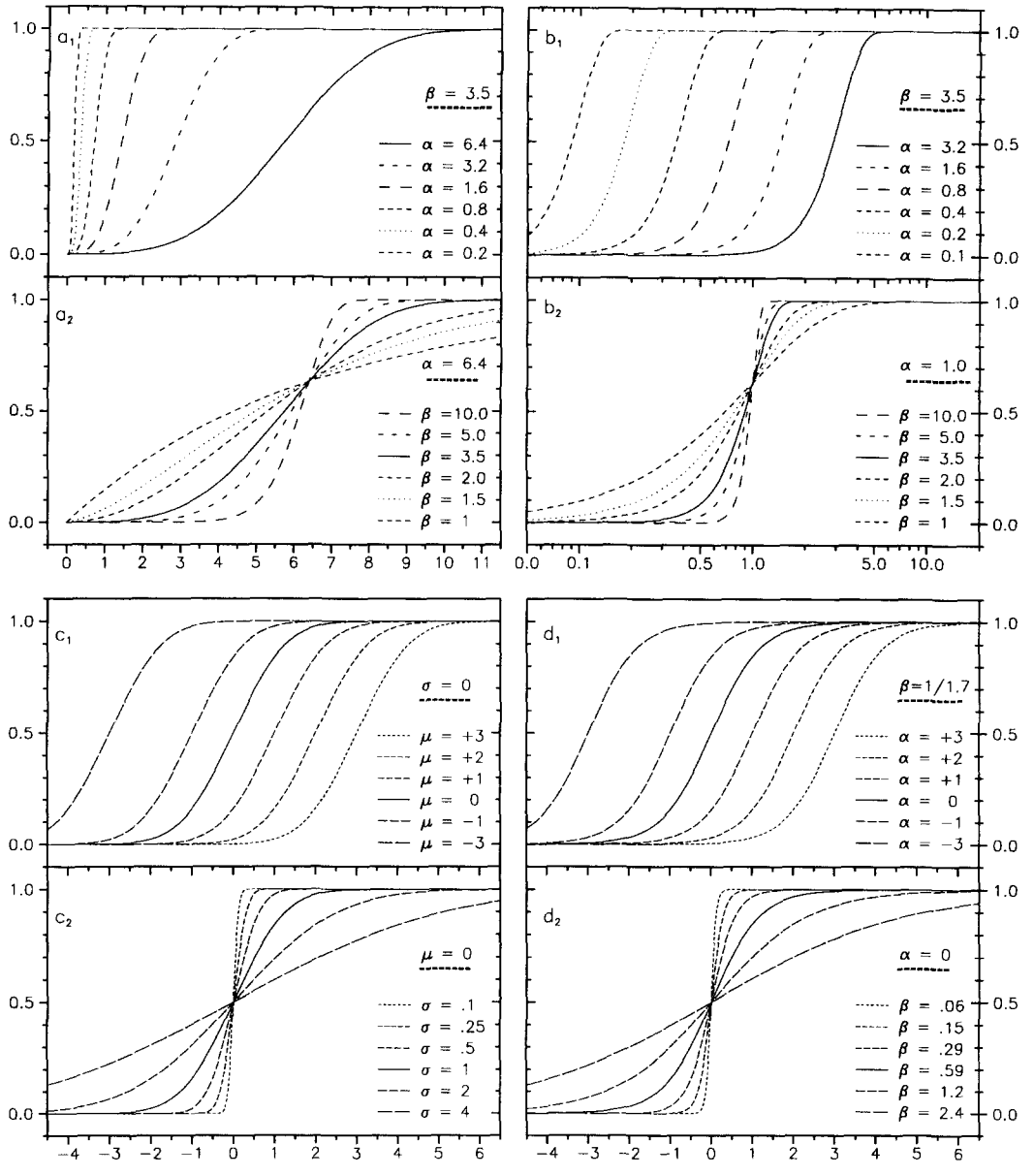

FIGURE 2. **Aspect of Psychometric Function Templates.** (a_1), (a_2) Cumulative Weibull distributions over a linear x-axis, (b_1), (b_2) cumulative Weibull distributions over a logarithmic x-axis, (c_1), (c_2) cumulative normal distributions (Gaussian distributions), and (d_1), (d_2) logistic functions. Each case is plotted for different values of the position parameter (index 1) and the slope parameter (index 2).

set is derived: first, new μ_{r+1}, σ_{r+1} are calculated according to the following formulas

$$\mu_{r+1} = \mu_r + (\hat{\mu}_r - \hat{\mu}_{r-1}) \frac{\hat{\mu}_{r-1} - \hat{\mu}_{r-2}}{\hat{\mu}_{r-1} + \hat{\mu}_{r-2}} \tag{18}$$

$$\sigma_{r+1} = \sigma_r + (\hat{\sigma}_r - \hat{\sigma}_{r-1}) \frac{\hat{\sigma}_{r-1} - \hat{\sigma}_{r-2}}{\hat{\sigma}_{r-1} + \sigma_{\text{adj}}} \tag{19}$$

$$\text{with} \quad \sigma_{\text{adj}} = \begin{cases} 0 & \text{if } \hat{\sigma}_{r-2} < \hat{\sigma}_{r-1} \\ \hat{\sigma}_{r-2} & \text{if } \hat{\sigma}_{r-2} > \hat{\sigma}_{r-1} \end{cases}$$

and second, applying equation (17) to these values of μ_{r+1}, σ_{r+1} yields four new values $x_{\{1..4\}}^{(r+1)}$ for the placement of the four constant stimuli. This means that the new stimulus set on the next block $r+1$ is not derived directly from the best estimates $\hat{\mu}_r, \hat{\sigma}_r$ after run r but by a kind of sliding estimates defined in equations 18 and 19. The fractional parts in equations 18 and 19 are therefore called by Watt and Andrews the *inertia* of the APE procedure. The inverse of the inertia is called *correction factor*. The inertia indicates that a sudden change in the subject's response behaviour, such as a shift of the threshold or a sequence of lapses, is not immediately reflected in the stimulus set. There is an asymmetry in the correction factor for σ given by σ_{adj} with the following reasoning. Usually, an experimental session starts with an overestimate of σ and therefore with a stimulus set which is too wide. Therefore, a decrease in the width of the stimulus set is more likely than an increase. The correction factors approach zero when the subject maintains a stable threshold. This fact could be used as a criterion for stopping the procedure but this was not noted by the authors. In personal communication Watt (1994) clarified that adapative pro-

A classical way to analyse probability data like that gathered in a psychophysical experiment, is the transformation of the data from the range $(0, 1)$ to the range $(-\infty, +\infty)$. A linear model is then adopted for the transformed value of the success probability. This procedure ensures that the fitted probabilities will lie between 0 and 1. Commonly used transformations are the logistic, the probit, and the complementary log–log transformation. They correspond to the logistic, the Gaussian (or normal) and the Gumbel distribution, respectively. An overview of these methods can be found in Collett (1991). In psychophysics the best known is probit analysis (Finney, 1971).

The probit transform assumes a cumulative normal distribution for the psychometric function. The probability data p is transformed into z-values by the inverse cumulative normal distribution (see Textbox 2 for the definition of $\mathcal{N}$)

$$p = \mathcal{N}(z)$$

i.e.

$$z(p) = \mathcal{N}^{-1}(p)$$

With the obtained z-values at the different stimulus levels x, parameters c_0 and c_1 are obtained by linear regression

$$z(x) = c_0 + c_1 x$$

and from

$$c_0 + c_1 x = \frac{x - \mu}{\sigma}$$

we derive

$$\mu = -\frac{c_0}{c_1} \quad \text{and} \quad \sigma = \frac{1}{c_1}.$$

Here μ is the value of the threshold and σ is the inverse of the slope of the psychometric function at the threshold.

The logistic transform, $\text{logit}(p)$, of a success probability p is given by the inverse of the logistic distribution (see Textbox 2)

$$\text{logit}(p) = \log \frac{p}{1 - p}$$

With the obtained logits we obtain c_0, c_1, α, β in the same way as before by linear regression, i.e.

$$\text{logit}(x) = c_0 + c_1 x = \frac{x - \alpha}{\beta}$$

results in $\alpha = -\frac{c_0}{c_1}$ and $\beta = \frac{1}{c_1}$.

The complementary log-log transform of a success probability p is given by the inverse of the Gumbel distribution (see also Textbox 2)

$$\text{co log}(p) = \log[-\log(1 - p)]$$

c_0, c_1, α, β are obtained in the same way as before, i.e.

$$\text{co log}(x) = c_0 + c_1 x = \frac{x - \alpha}{\beta}$$

results in $\alpha = -\frac{c_0}{c_1}$ and $\beta = \frac{1}{c_1}$.

Care should be taken that an iterative *weighted* linear regression is used. The weights depend on the number of trials at a given stimulus level and on the (unknown) probability of a correct answer; this is especially important if not all stimuli are presented equally often. When applied in psychophysics, the use of any of these transforms are problematic in the following two cases:

(1) When a small number of trials at each stimulus level is used: In this case it is quite likely that the responses at some levels are either all correct or all incorrect. All transformations are undefined for certainties (i.e. probability 1.0 or 0.0) and it is therefore not strictly possible to incorporate these data points in the linear regression, although they carry relevant information.

(2) For a non-zero guessing or lapsing rate p_g, p_l, which is accounted for by Abott's formula (equation 1 & 2), the weights have to be derived from the untransformed probabilities. It is possible to include p_g and p_l into the estimation but this would be again a nonlinear model (Collett, 1991, Chap. 4.4). For a forced choice design with two alternatives and a probit analysis of the results this problem has been pointed out by McKee, Klein and Teller (1985).

Textbox 3. Linearizing the psychometric function: Outline of probit, logit, and the complementary log-log transformation for converting cumulative probability data to a linear function.

bit estimation (APE) is run on a fixed number of trials basis, typically 64 blocks per session with 16 presentation each, totalling to 1024 trials for one experiment. The final estimates were derived from a probit analysis of all trials. He also noted that the current version of APE uses a sliding window (32 trials wide) to calculate the current stimulus set equation (17) after each trial. The final estimate is now calculated from the responses of the entire experiment via a maximum-likelihood technique.

Hall's hybrid procedure. Hall (1981) suggested to use a hybrid procedure, with the stimulus placement as given by PEST (Taylor & Creelman, 1967) in blocks of a predetermined number of presentations (Hall proposed 50). As a parametric model the logistic distribution is used. The first block starts with educated guesses of the threshold α_0 (position) and the spread β_0 (inverse slope) of the psychometric function. From the spread an initial step size δ for the PEST stepping rules is calculated by assigning $\delta = 4\beta_0$. After each block r, a maximum-likelihood estimation of *both* parameters of a logistic function (midpoint α_r and spread β_r) is performed by using a constrained gradient search method. * The constraints limit the search to the intervals $\alpha_r \in [\alpha_0 - \beta_0, \alpha + \beta_0]$ and $\beta_r \in [\frac{1}{c}\beta_0, c\beta_0]$ with $c = 10$. The new estimates α_r, β_r are used as initial values for the next $(r + 1)$ block of presentations. Every single PEST run starts with a stimulus value of $\alpha_r + 4\beta_r$ to give the subject a clear idea of what is to be detected.

Estimation of the threshold only

Particularly efficient methods of threshold estimation are obtained when the *slope* of the psychometric function is known in advance, i.e. only the location of the threshold is to be determined: The experimenter supplies not only the general form of the psychometric function but also its slope and other parameters (guessing and lapsing rate). This means that the different possibilities for the psychometric functions are translations of one *template* parallel to the x-axis (as shown by the examples in Fig. 2 (b_1,c_1, d_1)). The shape, i.e. most notably the fixed slope, of the psychometric function is predetermined by the experimenter. With one exception (the STEP method), the procedures in this section use either a Bayesian or a maximum likelihood estimator of the position parameter θ of the psychometric function $\psi(x;\Theta)$. Here Θ denotes the complete parameter set of the psychometric function whereas θ denotes the position parameter, i.e. the threshold. Examples of different psychometric function templates are shown in Textbox 2 (formulas) and Fig. 2 (aspect).

Statistical estimation theory. Estimation theory is a branch of probability theory and statistics that deals with the problem of deriving information about properties of random variables and stochastic processes for a given set of observed samples. A specific task is estimating a parameter of a population, given a set of data. There are four major construction principles of point estimators: minimum-χ^2, moments, maximum-likelihood, and Bayesian. In psychophysics maximum-likelihood and Bayesian estimators were proposed for the estimation of the threshold, given a set of binary responses. I therefore look at similarities and differences between them (see Textbox 4). Although both methods are very similar from the computational viewpoint, they differ considerably in their underlying assumptions and philosophical aspects (see e.g. Martz & Waller (1982, Chap. 5.1)).

Statistical decision theory. Statistical decision theory originated in the work of Wald (1947, 1950). It is concerned with the development of techniques for making decisions in situations where stochastic components play a crucial role. Important applications exist in business decision making, in operations research, and of course in psychophysics, where the problem is to decide at which stimulus level the next presentation should take place. The basic elements of statistical decision theory are:

(1) a space $\Omega_{\tilde{\theta}} = \{\tilde{\theta}\}$ which may be vector-valued, of the possible states of nature,
(2) an action space $A = \{a\}$ of the possible actions,
(3) a loss function $L(\tilde{\theta}, a)$ representing the loss incurred when action a is taken and the state of nature is $\tilde{\theta}$.

When estimating thresholds in psychophysics, the space $\Omega_{\tilde{\theta}}$ is the set of possible threshold values and the action space A is the set of presentable stimulus values. I don't know of any attempt at an explicit definition of a loss function for the psychophysical application, although King-Smith *et al.* (1994) note in this context that the mean of the posterior probability density function (pdf) minimizes the mean-squared error of the final estimate.

Application in psychophysics. The problem in psychophysics is two-fold: on the one hand the experimenter wants to estimate the threshold with the least possible number of trials, on the other hand he wants an optimal placement for these trials. The first problem is one of *sequential* estimation, first formulated by Wald (1947), and the latter is a problem of decision theory. When Bayesian or maximum-likelihood methods are used in psychophysics the goal is to use all available information † to place the next stimulus presentation as close as possible to the *true*, but unknown, location of the threshold. The best the experimenter can do is to place the stimulus presentation at the current best threshold estimate which is obtained by calculating the likelihood function, respectively the posterior probability density function (posterior pdf), sequentially during the experiment.

Calculation of the likelihood or unnormalized posterior pdf. After trial n of a session, when n stimulus presentations

*A similar way of fitting a cumulative Weibull distribution to psychometric data can be found in the appendix to Watson (1979).

†In the ML approach this is restricted to the information collected during the current experiment; whereas in the Bayesian approach general knowledge about the location of the threshold, e.g. the distribution of thresholds in some reasonable collective, can also be included.

Let $\mathbf{X}$ be a random variable whose probability density function (pdf) depends on some parameter set $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$. Let $f(\mathbf{x}|\Theta)$ denote the conditional joint pdf of one instance $\mathbf{x}$ of the random variable $\mathbf{X}$, which also depends on the parameter set Θ. The probability of obtaining exactly this instance $\mathbf{x}$ of $\mathbf{X}$ and corresponding responses $\mathbf{Z}$ is denoted by $f(\mathbf{x}|\Theta)$. In an actual experiment $f(\mathbf{x}|\Theta)$ can be calculated from the set of n presentations at different values of stimulus intensity $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and the corresponding responses $\mathbf{z} = (z_1, z_2, \ldots, z_n)$, if a parametric model for the psychometric function is assumed.

Maximum likelihood

The basic idea of the method of maximum likelihood is given by the following consideration: Different populations generate different data samples and any given data sample is more *likely* to have come from one population than from others. The method of maximum likelihood is based on the principle that we should estimate the parameter vector Θ, which describes the psychometric function, by its most plausible values, given the observed sample vector $\mathbf{x}$. In other words, the maximum likelihood estimators of $\theta_1, \theta_2, \ldots, \theta_k$ are those values of the parameter vector for which the conditional joint pdf $f(\mathbf{x}|\Theta)$ is at maximum. The name *likelihood function* denoted by $L(\Theta)$ is given to $f(\mathbf{x}|\Theta)$, viewed as a function of the parameter vector Θ. Therefore,

$$\mathbf{L}(\Theta) = f(\mathbf{x}|\Theta).$$

The maximum of the likelihood function is — as the name suggests — the maximum likelihood estimator for the parameter vector Θ.

Bayes' estimation

Important for the Bayesian viewpoint is (1) the concept of *subjective probability* and (2) not to look for a fixed value of a parameter, but to derive a probability distribution for the possible values of the parameters.

Let $g(\Theta)$ denote the prior pdf of Θ. Then, given the sample data vector $\mathbf{x}$, the posterior pdf $p(\Theta|x)$ of Θ is derived by Bayes' theorem

$$p(\Theta|x) = \frac{f(\mathbf{x}|\Theta)\, g(\Theta)}{h(\mathbf{x})}.$$

where $h(\mathbf{x})$ denotes a constant factor, depending only on the data $\mathbf{x}$, which is the normalizing marginal distribution of the posterior pdf $h(\mathbf{X}) = \int f(\mathbf{x}|\Theta)\, g(\Theta)\, d\Theta$. The unnormalized posterior pdf can be rewritten as

$$\begin{aligned}\mathcal{L}(\Theta) &= p(\Theta|x)h(\mathbf{X}) \\ &= f(\mathbf{x}|\Theta)g(\Theta).\end{aligned}$$

The function $\mathcal{L}(\Theta) = \mathbf{L}(\Theta)g(\Theta)$ is an unnormalized probability distribution for the parameter vector Θ.

It is easily seen that the only difference between the likelihood function and the posterior pdf is the multiplication by the prior pdf $g(\Theta)$. Apart from the philosphical aspects, the ML estimation is a special case of the more general Bayesian estimation. A Bayesian estimation with the mode (maximum) of the posterior pdf as estimator and a rectangular (uniform, or constant) *a priori* distribution is exactly equivalent to the ML-estimator. The rectangular prior pdf is called, in Bayesian terminology, a *non-informative prior*, or *a prior of ignorance*. The a priori distribution $g(\Theta)$ expresses our prior knowledge about the distribution of the parameter Θ, in which, e.g., the distribution of the thresholds in the population, or hardware constraints of the setup, can be incorporated. It is an example of a subjective probability (or a belief) about the parameter vector Θ.

Textbox 4. Comparison of maximum likelihood and Bayes estimation.

at intensities $(x_1, \ldots, x_n) = \mathbf{x}$ have taken place, the likelihood function (unnormalized conditional joint pdf for parameter vector Θ given the data $\mathbf{x}$, see also Textbox 4) is given by

$$\begin{aligned}\mathcal{L}(\Theta|x_1 \ldots x_n) = p(\Theta|x)f(\mathbf{x}) &= \prod_{i=1}^{n} \mathcal{L}(\Theta|x_i) \qquad (20)\\ &= \mathcal{L}(\Theta|x_n) \prod_{i=1}^{n-1} \mathcal{L}(\Theta|x_i)\end{aligned}$$

Each $\mathcal{L}(\Theta|x_i)$ is the probability that the subject has given a particular answer — correct or incorrect — at the stimulus intensity x_i which was presented at trial i. This probability is considered for different values of the parameter set Θ. The general formulation of the parameter set Θ stands for the multiparametric case, in the situation where only the single parameter of the threshold is estimated, this set Θ reduces to θ, the threshold.

In psychophysics, the probability of a particular answer is given by the psychometric function, i.e. $\psi(x_i, \theta)$. For the calculation of the likelihood the stimulus intensity x_i is fixed and the value of the threshold θ is considered as the variable. The likelihood function $\mathcal{L}(\theta|x_i)$ for a single trial i is given by

$$\mathcal{L}(\theta|x_i) = \begin{cases} p_+(x_i, \theta) = \psi(x_i, \theta) \\ p_-(x_i, \theta) = 1 - \psi(x_i, \theta) \end{cases} \qquad (21)$$

where p_+ and p_- stand for the probability for a correct (+) and incorrect (−) response of the subject. These probabilities are in turn given by the psychometric function and its complementary.

To facilitate the calculation of the likelihood, all versions have chosen the psychometric function to be translation invariant on the x-axis. The set of possible values

for the threshold θ_i coincides with the set of possible stimulus values x_i and the values of the single-trial likelihood can be calculated from a fast lookup-table.

In the case of ML estimation it was traditional to work with the logarithm of the likelihood rather than with the likelihood function itself. *

$$L_{\log}(\Theta) = \ln L(\Theta) = \ln f(\mathbf{X}|\Theta)$$

Use of the logarithm does not change the location of the mode (maximum), since the logarithm is a monotonic transform. However, its use complicates the calculation of both the mean and the median of the posterior pdf, and obscures the interpretation of the likelihood as a probability density for the location of the threshold when a specific data set is given. Although the distinction between the log-likelihood and unmodified probabilities sounds minor, it has caused some confusion about the interpretation of the log-likelihood and the posterior pdf in the psychophysical literature (Lieberman & Pentland, 1982; Emerson, 1986a).

A-priori density As pointed out in Textbox 4, the a-priori density $g(\Theta)$ expresses our prior knowledge about the distribution of the parameter Θ, in which, for example, the distribution of the thresholds in the population, or hardware constraints of the setup, can be incorporated. In equation (20) the history of the session up to trial $n-1$ is represented by the term $\prod_{i=1}^{n-1} \mathcal{L}(\Theta|x_i)$. This term expresses our posterior knowledge at trial n, but can also be looked at as the prior pdf for trial $n+1$. It includes to some degree the information contained in the prior density at the beginning of the experiment. If, at the start of a session, the experimenter has no information about where the threshold could be, a simple solution is a rectangular, or uniform prior density, which assigns every possible value of the threshold location the same probability. Since in almost all practical cases the experimenter has some knowledge about the location, this information should be included, e.g. by using a stepwise uniform density having a higher probability in some subrange than in the rest of the interval. The experimenter should ensure that the prior pdf at the beginning never dominates the posterior pdf at the end of the experiment (see Martz & Waller, 1982, last paragraph of Sec. 5.2 for a discussion on dominant likelihoods). This is normally the case when the prior density is relatively flat compared to final posterior pdf. The use of a deliberately choosen prior density can speed up the experiment since it favours presentations in the beginning of the experiment at reasonable values. An example where bad stimulus placement during the first few trials can be counterbalanced by a reasonable prior, is the behaviour of pure maximum likelihood methods (uniform prior and the maximum of the posterior pdf as estimator): If the first response of the subject is correct, the second trial will be presented at the lowest, and if it is incorrect, at the highest possible value. When a prior is used as, e.g. in ZEST, QUEST and ML-TEST, these large jumps into uninteresting regions during the first trials are reduced.

A-posteriori density. King-Smith *et al.* (1994) have distinguished, as mentioned above, *measurement* bias from *interpretation* bias. The latter they relate to the *a-posteriori* probability density for the location of the threshold. The value of the posterior pdf at a given stimulus level is the probability † that the current set of answers are obtained with the threshold at this level. In the discrete case, the posterior pdf is given by

$$p_{\text{post}}(\theta, \mathbf{x}) = \frac{\mathcal{L}(\theta|x_1 \ldots x_n)}{\sum_{j=1}^{m} \mathcal{L}(\theta_j|x_1 \ldots x_n)}, \qquad (22)$$

where n is the number of trials done, and m indexes the number of different stimulus values which are under consideration for being the threshold. It is important for the correct interpretation of the posterior pdf that it brackets the threshold sufficiently, which means that the probability for the threshold lying at either of both ends is neglectable.

Best estimate. The current best estimate of the threshold is — in the ML approach — the location of the maximum (mode) of equation (20). In Bayesian theory there is no single best estimate. The estimator is different for different loss functions. If a squared-error loss function is used, the best estimate is given by the mean of the posterior pdf. For an absolute-error loss function the median of the pdf is the best estimator. The mode (maximum) of the posterior pdf has the intuitively appealing interpretation of being the "most likely" or plausible value, given the prior and the data, but cannot be derived from any of the standard loss functions. For the application in psychophysics, King-Smith *et al.* (1994) and Emerson (1986a) were able to show by simulations that the mean is superior to the median or the mode: With the same number of trials, the mean yielded more reliable and less biased estimates of the psychophysical threshold than the mode.

Pelli (1987a, b) suggested using the value as an estimator that minimizes the variance of the posterior pdf by looking ahead and calculating all possible combinations till the end of the experiment. King-Smith *et al.* (1994) have shown that with a one and two trial look-ahead this minimum variance estimator has only slight advantages over the mean. Since the mean of the posterior pdf as estimator minimizes the variance of the estimate, the computational overhead necessary for doing the look-ahead can be dispensed with.

Termination rules. For the termination of the sequential estimation procedure most of the implementations advise

*This was more convenient to work with in the times of slide rules and hand calculation, and for analytical solutions. With the powerful personal computers available nowadays, the computational time saving is irrelevant.

† Sadly, this is completely true only if the psychometric function used in equation (21) matches that of the subject in *all* degrees of freedom.

the experimenter to use a fixed number of trials, although Dantzig (1940), as cited by Sen (1985), has shown that the estimation of the location parameter θ within a prespecified confidence interval (θ_l, θ_u) is impossible with a fixed sample size when the variance of the underlying distribution is unknown. For psychophysical thresholds, where the slope of the psychometric function is not exactly known in advance, this implies that it is impossible to obtain threshold estimates with a predetermined variance when only a fixed number of trials are performed. Therefore it is preferable to use a dynamic termination criterion within the framework of sequential statistics. A session is then ended when a desired level of confidence in the obtained threshold location is reached, i.e. when a predetermined variance of the threshold estimate is attained.

In the Bayesian approach the posterior pdf is a probability density for the location of the threshold parameter. Therefore, for a desired confidence level of γ, solving the following equations leads to a condition where the best estimate of the threshold value lies in the interval (θ_l, θ_u) with probability γ

$$\int_{-\infty}^{\theta_l} p(\theta|\mathbf{x})d\theta = \frac{1-\gamma}{2} \quad \text{and} \quad \int_{\theta_u}^{+\infty} p(\theta|\mathbf{x})d\theta = \frac{1-\gamma}{2}. \quad (23)$$

In most cases equation (23) can easily be calculated from equation (22), the posterior pdf or *normalized* likelihood by summing up the values of the posterior pdf from the best estimate in both directions until a cumulative value of $\frac{\gamma}{2}$ is reached. The upper and lower bounds are the stimulus values which correspond to these two values.

In a different approach, Watson and Pelli (1983) and Harvey (1986), referring back to Wilks (1962), advised a likelihood-ratio test for determining a confidence interval. The arguments given by these authors for applying the test are not fully developed and three questions arise: first, the test does not account for the sequential nature of the estimation problem. Second, it is only asymptotically valid, i.e. only for a large number of trials. Third, it requires the underlying function to be a cumulative distribution function, which is not the case for the psychometric function as explained in equation (3). The further development of QUEST by Laming and Marsh (1988) and their approximation to the variance of the best estimate provides a better solution. The escape from parametric models as given by Sen (1985), especially the derivation of sequential nonparametric confidence intervals (Chap. 4.5) and the sequential likelihood-ratio test (SLRT, Chap. 6.2) might lead to a new solution to the termination problem.

The tests which are currently used for termination have a slight drawback due to their parametric nature: The width of the posterior pdf is influenced by the slope of the psychometric function, which is used to calculate the posterior pdf [see equation (21)]. This conforms to my experience with the Bayesian probability interval approach and, according to Madigan and Williams (1987), also for the likelihood ratio approach mentioned above. As a result, for an assumed steep slope the posterior pdf tends to be very narrow and the confidence intervals tend to be too small, and therefore the resulting sessions tend to be very short. For shallow slopes the posterior pdf is broad and confidence intervals are too large and the resulting sessions tend to be very long. As for any parametric model, correct confidence/probability intervals are only obtained if the assumed slope matches that of the subject. To be on the safe side with such a dynamic stopping criterion, it is advisable to underestimate the slope (which is equivalent to an overestimation of the spread, or standard deviation) used in equation (21) relative to the subject's slope. This guarantees that the *true* confidence interval is narrower than the one which is calculated from the posterior pdf and the sessions tend to be slightly longer than it would be necessary. From preliminary results of extensive simulations* I can provide the following rule of thumb. Adjust one of the following three parameters, the slope, the width of the confidence interval, or the confidence level, until the procedure stops after the following number of trials, on average: yes–no method, 20; 8-AFC, 25; 4-AFC, 30; 3-AFC, 37.5; 2-AFC, 50.

Psychophysical incarnations of Bayesian and ML-methods. Since the first intimation of a sequential maximum-likelihood procedure by Hall (1968, 1981) several others have been suggested: QUEST (Watson & Pelli, 1979, 1983) BEST PEST (Pentland, 1980; Lieberman & Pentland, 1982), ML-TEST (Harvey, 1986), QUADRATURE METHOD (Emerson, 1986b), IDEAL (Pelli, 1987a, b), YAAP (Treutwein, 1989, 1991; Treutwein & Rentschler, 1992), and ZEST (King-Smith, Grisby, Vingrys, Benes & Supowit, 1991; King-Smith *et al.*, 1994). All use maximum-likelihood or Bayesian estimators and differ only in minor aspects (for an overview see Table 2). ML-TEST, QUEST and ZEST use a psychometric function defined over the physical domain of the stimulus intensity variable, which is scaled in logarithmic steps. These methods therefore use a cumulative Weibull distribution, although ML-TEST leaves it up to the user to choose from three different psychometric functions (Weibull, cumulative normal and logistic). BEST PEST and YAAP implement a logistic function, which is defined over the index of the likelihood array. ML-TEST and QUEST use heuristic priors, where ML-TEST simulates a small number of trials at a guessed threshold location and QUEST fills the prior with a broad normal density centred at a guessed location of the threshold. Both, ML-TEST and QUEST do not include this prior for calculating the final estimate. King-Smith *et al.* (1994) have chosen for their implementation of ZEST to use the following prior: the prior is calculated from an analytical approximation to the distribution of the threshold in a representative group† of normal subjects and patients. They also checked that this prior does

*I am currently working on a comparitive evaluation of adaptive psychophysical procedures, the influence of different parameters and their mismatches on the estimate.

†For two different tasks a total number of 18,944 and 70,247 thresholds were included in these histograms.

TABLE 2. Comparison of important details of Bayes' and maximum likelihood adaptive procedures

		Estimator	Prior pdf	Termination criterion
Hall	(1968)	ML – mode	uniform	number of trials
QUEST	(1979)	Bayes – mode	normal density with mean and variance specified by experimenter	number of trials
BEST PEST	(1980)	ML – mode	uniform	number of trials
ML-TEST	(1986)	ML – mode	small number of simulated trials at a specified threshold	χ^2 test
Emerson	(1986)	Bayes – mean	uniform	number of trials
IDEAL	(1987)	Bayes – minimum variance lookahead	normal density with mean and variance specified by experimenter	number of trials
YAAP	(1989)	Bayes – mode/mean	uniform	probability interval
ZEST	(1991)	Bayes – mean	analytic approximation to test group histogram	number of trials

not dominate the posterior. All other procedures implement a rectangular, or — in Bayesian terminology — noninformative, prior. Only YAAP and ML-TEST implement a dynamic stopping criterion; ML-TEST uses the likelihood-ratio approach and YAAP uses the probability interval. YAAP and ZEST use unmodified probabilities, all other methods work with the log-likelihood. ZEST, YAAP * and Emerson's procedure use the mean of the a posteriori for placing the stimuli and calculating the final estimate, all other procedures implement the mode.

The STEP Method is a special case: Simpson (1989) suggested to use a step function for the psychometric function and to fit the step function to the binary single trial responses. Although Simpson claims that one of the main advantages of his method is that no slope has to be specified, the use of a step function is equivalent to using a standard psychometric function template (Textbox 2 and Fig. 2) with a slope of infinity. Moreover, a step function is discontinuous and all standard regression routines, linear or nonlinear, require continous and differentiable functions. The derivation of his algorithm is therefore highly questionable and is not backed up by adequate references †. It is possible that the STEP method can be explained in the concepts of a maximum-likelihood method or of a linear regression to a transformed psychometric function (see Textbox 3; in the latter case a special sequential linear regression using indicator variables, see Collett, 1991, Chap. 3.2.1) and thus given a sounder basis. In both cases, however, his claim that no slope for the psychometric function must be specified, is off the point. Simpson's reasoning is mainly *ad hoc* and based on Monte-Carlo simulations, which were criticized by Watson and Fitzhugh (1990) in that they do not constitute a reasonable model of the experimental process.

*After I received a preprint of King-Smith *et al.* (1994) in 1993, I switched from mode to mean.

†In personal communication Simpson refered to linear regression using indicator variables (Neter, Wassermann and Kutner (1990), Chap. 10).

EVALUATION

Adaptive procedures can be compared by conducting threshold measurements with real subjects or by constructing computer simulations. Both methods of evaluation present different problems:

- When real experiments with subjects are used to evaluate psychophysical methods, as done, e.g. by Hesse (1986), McKee *et al.* (1985), O'Regan and Humbert (1989), Shelton and Scarrow (1984), Stillmann (1989) or Woods and Thomson (1993), the variability of the *estimated* thresholds is overloaded by:
  - the variability of the *true* threshold in subjects, be it inter- or intraindividual variability, e.g. circadian variations, attention, alertness, or sleepiness, and
  - systematic trends of the true threshold, e.g. learning, masking, or adaptation.

  It is very difficult to separate these effects in real experiments from systematic or random errors in the estimation process, i.e. to decide whether the variability is due to bias or insufficient precision of estimation process or to true variability in the subjects.
- When simulations are used to evaluate psychophysical methods, as done, e.g. by Emerson (1986a), Kershaw (1985), Leek, Hanna and Marshall (1992), Lieberman and Pentland (1982), Madigan and Williams (1987), Maloney (1990), Rose, Teller and Rendleman (1970), Simpson (1989) or Swanson and Birch (1992). In this case, a model for the psychophysical observer is chosen and responses are generated according to the probabilities associated with this observer model. The following questions arise:
  - Does the chosen observer model reflect a *real* observer's behaviour? e.g. does it include guessing and/or lapsing rates [see equations (1) and (2)]? What happens, if one of these parameters or the slope of the simulated

observer does not match the parameters of the model? This is especially important for parametric methods like APE, Hall's hybrid method and all Bayesian/ML methods.

- What happens, and how can one simulate an observer in a forced choice design who is biased towards one interval? Human observers are known to behave non-randomly; if they are asked to produce random numbers, they avoid sequential repetitions (Brugger, Landis & Regard, 1990). How can this behaviour be included in an observer model?
- Does the random number generator fulfil what it promises * ?

The only computer simulation study which was able to overcome the problem of random number generation is by King-Smith *et al.* (1994) where different Bayesian estimators were evaluated by enumerating *all* combinations of possible response sequences. One of the important result of this study was that the mean of the posterior pdf gives unbiased estimates.

A complete comparative evaluation — even of a representative subset only — is beyond the scope of this article. There are too many points, which have to be taken into consideration. To extract the evaluation from published material does not really help since many of the evaluative studies do not cover a representative subset of adaptive procedures. Some ignore complete groups of methods, some of the evaluations seem to be biased toward the procedure suggested by the author. Furthermore there is only rare published material on the important matter of sequential dependencies, e.g. learning, masking, or adaptation, which violate the requirement of stationarity. The stationarity is a minimum requirement for all adaptive methods. But also in the classical methods non-stationarity influences the results; in the method of constant stimuli, e.g. non-stationarity is reflected by a shallower psychometric function.

The influence of different levels of the lapsing rate was investigated by Swanson and Birch (1992). They compared maximum-likelihood estimates with UDTR estimates for four levels of p_l (0%, 5%, 10%, and 15%) and found unacceptable bias for the ML-methods introduced by high lapsing rates. In their study, the parametric model did not reflect the lapsing rates, but this is similar to a real experiment where the experimenter does not normally know the guessing or lapsing rates before the experiment. One of the consequences could be to include a certain number of "catch trials", randomly interleaved in the normal sequence, which are located far above or below the estimated threshold, to simultaneously estimate these rates. The results of this estimation should be used in the parametric model for the final estimate which means a reanalysis of the experiment after it's end with a more correct model.

If two parameters are estimated, i.e. threshold and slope of the psychometric function, O'Regan and Humbert (1989) found that small samples (100 data points simulated with $n = 10$ number of presentations at $N = 10$ stimulus values, method of constant stimuli) produce both, low precision and biased estimates. These results were obtained for either maximum-likelihood and probit analysis and they are in accordance with the study of McKee *et al.* (1985). Similar results were found by Leek *et al.* (1992) who compared the method of constant stimuli, APE, and UDTR.

Madigan and Williams (1987) compared QUEST, BEST PEST, and PEST in a yes–no and a two-alternative forced choice situation with slope mismatches. They found that moderate mismatches — apart from the influence on a dynamic termination criterion — do not produce adverse effects on the estimation process. Similar results were found by Green (1990). According to my own experience shallow slopes tend to distribute the presentations more widely around the "true" threshold value whereas steep slopes focus the presentations around that value. At the same time, steep slopes make procedures very susceptible to mismatched guessing and lapsing rates and easily yield outlying "bad" estimates, whereas shallow slopes tend to tolerate these mismatches between the assumed parametric model and the subject's "true" parameters easier.

DISCUSSION

After almost a decade, there is still no solution to the problems pointed to by Harvey (1986, p. 629f):

> "Subjects in psychophysical experiments do violate both the assumption of stationarity and of independence (Shipley, 1961; Taylor, Forbes & Creelman, 1983) by showing lapses and sequential dependencies of their responses. [...] When psychometric functions are measured in experiments in which the subject is asked to report wether or not the stimulus was seen (or heard), for example, the dependent variable is the hit rate. Several authors (Nachmias, 1981; Watson & Pelli, 1983) have suggested that these data may be fit with logistic or Weibull functions setting γ [the guessing rate] equal to the false alarm rate. The problem with this approach is that the false alarm rate is not constant: there is a different false alarm rate for each point on the psychometric function."

Harvey suggested forced choice methods as a solution for these problems. For being criterion free, forced choice requires an unbiased observer, which cannot be taken for granted. Besides the problem of non-random behaviour of subjects, it is possible that the observer in spatial forced choice designs, has a preferred location for guessing. In temporal forced choice designs interactions in time are possible. Masking or afterimages, e.g. are quite pronounced at low spatial frequencies and can yield to a preference of one of the intervals. Forced choice methods can also lead to practical difficulties, e.g. in a clinical setting, when a patient says "You are forcing me to guess and you want to base your final diagnosis on what I guess. Is this really a trustworthy method?" It is impossible to explain to most of these patients the

*This is a point which is easily overlooked. Low quality library routines for generating random numbers are frequently used. Some of them produce numbers having serial correlations or short term cycles, see Park and Miller (1988), and Press *et al.* (1992).

benefits of forced choice designs like the independence of the criterion, etc. One solution could be a kind of "signal detection" adaptive testing, where the observer is required to assign his same/different judgement to several categories and the experimenter evaluates these categorial responses dynamically in terms of signal detection or choice theory. This means evaluating sensitivity (e.g. d', α) and criterion (e.g. c, b) on line and placing the stimulus presentations according to the sensitivity results. Another important drawback of forced choice experiments, when the number of alternatives is is limited to two, is the fact that the necessary number of trials to reach a certain precision must be increased at least by a factor of 2 – 3 compared to a yes–no design (see McKee *et al.*, 1985; King-Smith *et al.*, 1994).

To avoid sequential interactions, it is a good idea to interleave independent runs for different parameters in one session. For example, in an experiment for measuring contrast sensitivity, the runs for the different spatial frequencies could be randomly interleaved and thereby sequential interactions between succeeding trials are minimized.

CONCLUSIONS

Although the methods of stochastic approximation were advocated for sensitivity data 43 years ago and mentioned subsequently for psychophysical application by several other authors (Smith, 1961; Falmagne, 1986; King-Smith *et al.*, 1994), they were ignored by practising psychophysicists. It is time that these methods should receive more attention. Since they are non-parametric, as opposed to the Bayesian and maximum-likelihood methods, it can be expected that psychophysical procedures based on stochastic approximation are less susceptible to parameter mismatches than are parametric methods. Preliminary results of simulations indicate that the accelerated stochastic approximation has a similar performance as mean-Bayesian methods, which seems to be near optimal performance. Besides other valuable theoretical results (for overviews see Dupač, 1984; Sampson, 1988), Taylor (1971) has used the asymptotic variance of the Robbins-Monro process, i.e. stochastic approximation, as the touchstone for evaluating performance of adaptive procedures. To my knowledge no one has applied any of these methods directly to psychophysics.

Until there are usable results concerning the performance of stochastic approximation methods, I advise the following, depending on what the experimenter wants to know:

- Threshold and slope: there is no clear winner here. Probably a method like APE could be further developed in the following way: keep two stimuli far from threshold, one well above and one well below, for estimating the guessing and lapsing rate and then target three stimulus sequences (e.g. by a stochastic acceleration method) at values reasonably spaced between the guessing and lapsing rate i.e. one at the percentage correct for the threshold and the other two two-thirds of the difference between the guessing/lapsing rate and the threshold performance above and below the threshold performance. Estimating could be done by some appropriate method, like a linearized approach, a maximum likelihood estimator, a two-dimensional Bayesian mean estimator, or a nonlinear regression routine like the Levenberg-Marquardt compromise. *
- Only the threshold: use a Bayesian method with a mean estimator and a dynamic termination criterion, which terminates after about 20 trials for yes–no, about 50 trials for a two-alternative forced choice design, or about 30–25 trials for a 4–8 alternative design.

REFERENCES

Bebie, H., Fankhauser, F., & Spahr, J. (1976). Static perimetry: Strategies. *Acta Ophthalmologica, 54*, 325–338.

Blum, J. R. (1954). Approximation methods which converge with probability one. *Annals of Mathematical Statistics, 25*, 382–386.

Brugger, P., Landis, T., & Regard, M. (1990). A "sheep–goat effect" in repetition avoidance: Extra sensory perception as an effect of subjective probability? *British Journal of Psychology, 81*, 455–468.

Collett, D. (1991). *Modelling binary data*. London: Chapman & Hall.

Daintith, J. & Nelson, R. D. (1989). *Dictionary of mathematics*. London: Penguin Books.

Dantzig, G. B. (1940). On the non-existence of tests of 'students' hypothesis' having power function independent of σ. *Annals of Mathematical Statistics, 11*, 186–192.

Derman, C. (1957). Non-parametric up-and-down experimentation. *Annals of Mathematical Statistics, 28*, 795–797.

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

Dupač, V. (1984). Stochastic approximation. In Krishnaiah, P. R. & Sen, P. K., eds, *Handbook of statistics*, volume 4, pp. 515–529. Elsevier.

Emerson, P. L. (1986a). Observations on maximum likelihood and Bayesian methods of forced choice sequential threshold estimation. *Perception & Psychophysics, 39*, 151–153.

Emerson, P. L. (1986b). A quadrature method for bayesian sequential threshold estimation. *Perception & Psychophysics, 39*, 381–383.

Falmagne, J. C. (1986). Psychophysical measurement and theory. In Boff, K. R., Kaufman, L., & Thomas, J. P., eds, *Handbook of perception and human performance*. New York: John Wiley & Sons.

Fechner, G. T. (1860). Elemente der Psychophysik. English translation: Howes, D. H. & Boring, E. C. (eds) and Adler, H. E. (transl.), New York: Holt (Rinehart & 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*. 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. & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Los Altos, Calif.: Peninsula. Reprinted 1988.

Hall, J. L. (1968). Maximum-likelihood sequential procedure for estimation of psychometric functions. *Journal of the Acoustical Society of America, 44*, 370.

*This method is an elegant compromise between a steepest descent minimization and a global search for the minimum. Good descriptions of the Levenberg-Marquardt method are Press *et al.* (1992, Chap. 15.5), or, in full detail, Reich (1992).

Hall, J. L. (1981). Hybrid adaptive procedure for estimation of psychometric functions. *Journal of the Acoustical Society of America, 69*, 1763–1769.

Harvey, jr, L. O. (1986). Efficient estimation of sensory thresholds. *Behavior Research Methods, Instruments & Computers, 18*, 623–632.

Hesse, A. (1986). Comparison of several psychophysical procedures with respect to threshold estimates, reproducibility and efficiency. *Acustica, 59*, 263–273.

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

Kaplan, H. L. (1975). The five distractors experiment: Exploring the critical band with contaminated white noise. *Journal of the Acoustical Society of America, 58*, 504–511.

Kershaw, C. D. (1985). Statistical properties of staircase estimates from two interval forced choice experiments. *British Journal of Mathematical and Statistical Psychology, 38*, 35–43.

Kesten, H. (1958). Accelerated stochastic approximation. *Annals of Mathematical Statistics, 29*, 41–59.

King-Smith, P. E., Grisby, S. S., Vingrys, A. J., Benes, S. C., & Supowit, A. (1991). Evaluation of four different variations of the QUEST procedure for measuring thresholds. *Investigative Ophthalmology and Visual Science (Suppl.), 32*, 1267.

King-Smith, P. E., Grisby, S. S., Vingrys, A. J., Benes, S. C., & Supowit, A. (1994). Comparison of the QUEST and related methods for measuring thresholds: Efficiency, bias and practical considerations. *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. (1970). Transformed up–down methods in psychoacoustics. *Journal of the Acoustical Society of America, 33*, 467–476.

Lieberman, H. R. & Pentland, A. P. (1982). Microcomputer-based estimation of psychophysical thresholds: The best PEST. *Behavior Research Methods, Instruments & Computers, 14*, 21–25.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

Luce, R. D. (1963). Detection and recognition. In Luce, R. D., Bush, R. R., & Galanter, E., eds, *Handbook of mathematical psychology*, volume 1, pp. 103–189. New York: Wiley.

MacAdam, D. L. (1942). Visual sensitivities to color differences in daylight. *Journal of the Optical Society of America, 32*, 247–274.

Macmillan, N. A. & Creelman, D. C. (1991). *Detection theory: a users's guide*. Cambridge: Cambridge University Press.

Madigan, R. & Williams, D. (1987). Maximum-likelihood psychometric procedures in two-alternative forced-choice: Evaluation and recommendations. *Perception & Psychophysics, 42*, 240–249.

Maloney, L. T. (1990). Confidence intervals for the parameters of psychometric functions. *Perception & Psychophysics, 47*, 127–134.

Martz, H. F. & Waller, R. A. (1982). *Bayesian reliability analysis*. New York: John Wiley & Sons.

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.

Nachmias, J. (1981). On the psychometric function for contrast detection. *Vision Research, 21*, 215–223.

Neter, J., Wassermann, W., & Kutner, M. H. (1990). *Applied linear statistical models*. Boston: Irwin.

O'Regan, J. K. & Humbert, R. (1989). Estimating psychometric functions in forced-choice situations: Significant biases found in threshold and slope estimations when small samples are used. *Perception & Psychophysics, 45*, 434–442.

Park, S. K. & Miller, K. W. (1988). Random number generators, good ones are hard to find. *Communications of the ACM, 31*, 1192–1201.

Pelli, D. G. (1987a). The ideal psychometric procedure. *Investigative Ophthalmology and Visual Science (Suppl.), 28*, 336.

Pelli, D. G. (1987b). The ideal psychometric procedure. *Perception (Suppl.), 16*, 237.

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

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in C: the art of scientific computing (2nd edn)*. Cambridge: Cambridge University Press.

Reich, J. G. (1992). *C Curve fitting and modelling for scientists and engineers*. New York: McGraw-Hill.

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.

Sampson, A. R. (1988). Stochastic approximation. In Kotz, S. & Johnson, N. L., eds, *Encyclopedia of statistical sciences*. New York: Wiley.

Sen, P. K. (1985). *Theory and applications of sequential nonparametrics*. Philadelphia: SIAM.

Shelton, B. R. & Scarrow, I. (1984). Two-alternative versus three-alternative procedures for threshold estimation. *Perception & Psychophysics, 35*, 385–392.

Shipley, E. (1961). Dependence of successive judgments in detection tasks: Correctness of response. *Journal of the Acoustical Society of America, 71*, 1527–1533.

Silberstein, L. & MacAdam, D. L. (1945). The distribution of color matchings around a color center. *Journal of the Optical Society of America, 35*, 32–39.

Simpson, W. A. (1989). The STEP method: A new adaptive psychophysical procedure. *Perception & Psychophysics, 45*, 572–576.

Smith, K. J. E. (1961). Stimulus programming in psychophysics. *Psychometrika, 26*, 27–33.

Spahr, J. (1975). Optimization of the presentation pattern in automated static perimetry. *Vision Research, 15*, 1275–1281.

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

Swanson, W. H. & Birch, E. E. (1992). Extracting thresholds from noisy psychophysical data. *Perception & Psychophysics, 51*, 409–422.

Taylor, M. M. (1971). On the efficiency of psychophysical measurement. *Journal of the Acoustical Society of America, 49*, 505–508.

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. (1989). Adaptive psychophysical procedures. *Perception (Suppl.), 18*, 554.

Treutwein, B. (1991). Adaptive psychophysical methods. In Bhatkar, V. P. & Rege, K. M., eds, *Frontiers in knowledge-based computing*. New Delhi: Narosa.

Treutwein, B. & Rentschler, I. (1992). Double pulse resolution in the visual field: The influence of temporal stimulus characteristics. *Clinical Vision Sciences, 7*, 421–434.

Treutwein, B., Rentschler, I., & Caelli, T. M. (1989). Perceptual spatial frequency-orientation surface: Psychophysics and line element theory. *Biological Cybernetics, 60*, 285–295.

Tyrell, R. A. & Owens, D. A. (1988). A rapid technique to assess the resting states of eyes and other threshold phenomena: The modified binary search (MOBS). *Behavior Research Methods, Instruments, & Computers, 20*, 137–141.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

Wald, A. (1950). *Statistical decision functions*. New York: Wiley.

Watson, A. B. (1979). Probability summation over time. *Vision Research, 19*, 515–522.

Watson, A. B. & Fitzhugh, A. (1990). The method of constant stimuli is inefficient. *Perception & Psychophysics, 47*, 87–91.

Watson, A. B. & Pelli, D. G. (1979). The QUEST staircase procedure. *Applied Vision Association Newsletter, 14*, 6–7.

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.

Wilks, S. (1962). *Mathematical statistics*. New York: Wiley.
Woods, R. L. & Thomson, W. D. (1993). A comparison of psychometric methods for measuring the contrast sensitivity of experienced observers. *Clinical Vision Sciences, 8*, 401–415.

Acknowledgements — This study has been supported by the Deutsche Forschungsgemeinschaft grants Po 131/13 and Re 337/7. I would like to thank Terry Caelli, Mario Ferraro, Patrick Flannagan, David Foster, Lewis O. Harvey jr, Robyn Hudson, Martin Jüttner, Ewen King-Smith, Ingo Rentschler, Hans Strasburger and Christoph Zetzsche for helpful comments and/or enlightening discussions during the years of struggle with the topic of adaptive psychophysical procedures and/or draft versions of this manuscript. I am also grateful to the reviewers Denis Pelli and Roger Watt as well as the special editor Michael Morgan for encouraging reviews and comments on the first version of the manuscript, which helped to enhance it significantly. Special thanks go to Alan Cowey who read the last version carefully and made many final improvements.