Perception & Psychophysics
1999, 61 (1), 87-106

Fitting the psychometric function

BERNHARD TREUTWEIN
Ludwig-Maximilians-Universität, Munich, Germany

and

HANS STRASBURGER
Otto-von-Guericke-Universität, Magdeburg, Germany

A constrained generalized maximum likelihood routine for fitting psychometric functions is proposed, which determines optimum values for the complete parameter set—that is, threshold and slope—as well as for guessing and lapsing probability. The constraints are realized by Bayesian prior distributions for each of these parameters. The fit itself results from maximizing the posterior distribution of the parameter values by a multidimensional simplex method. We present results from extensive Monte Carlo simulations by which we can approximate bias and variability of the estimated parameters of simulated psychometric functions. Furthermore, we have tested the routine with data gathered in real sessions of psychophysical experimenting.

The psychometric function—an analytic function that relates the proportion of correct responses in a sensory task to some physical stimulus value—plays a basic role in psychophysics. Independent of the question of whether a "true" sensory threshold exists and what its nature may be, the psychometric function permits a concise description of empirical data and allows predictions about sensory performance. For the probabilistic description of performance in terms of a response continuum, Corso (1963) used the term *response threshold*, which is operationally defined as a specific point on the psychometric function. In recent years, the question of how the psychometric function can be determined has regained interest through the development of improved, adaptive threshold measurement techniques (for an overview, see Treutwein, 1995), of which those using maximum likelihood (ML) or Bayesian principles seem particularly promising. These methods owe their efficiency to the fact that they make assumptions about the psychometric function's shape, expressed by its parameters, although these assumptions have rarely been checked.

The psychometric function can be fully described by four parameters: one defining the function's position on the abscissa (often referred to as the *threshold*), one defining its *slope* or *spread* (the inverse of slope), and one each for defining the upper and the lower asymptote (extended version of *Abbott's formula*; Finney, 1971, Chap. 7.1; Macmillan & Creelman, 1991, Chap. 5.4.2; see also Equation 3 and Figure 1). Of the two asymptotes, the lower asymptote describes the performance at low stimulus levels. In a forced-choice design, this performance is obviously governed by guessing, and the parameter for the low asymptote is, therefore, often called the *guessing rate*.[1] In a yes/no design, the performance at low stimulus levels is attributed to intrinsic noise. A potential difficulty with that name is that it implies that guessing itself is independent of the stimulus value;[2] although this independence has been disproved by several authors (e.g., Green & Swets, 1966; Nachmias, 1981), we wish to adhere to this term, since it serves as a convenient description of a subject's performance at low stimulus intensities. The upper asymptote describes performance at high stimulus levels at which the sensory mechanism is generally assumed to be perfect; misses are attributed to failure of the equipment or to attentional lapses of the subject. These deviations from perfectness are, therefore, commonly called the *lapsing rate* (see note 1).

In practical applications of psychometric function fitting—for example, in most adaptive procedures (see Treutwein, 1995)—only one or, rarely, two parameters are estimated, usually those of threshold and slope (see, e.g., Hall, 1968; Watson, 1979) or, sometimes, threshold and lower asymptote (Green, 1993). The parameters that are not estimated are set to some reasonable value: the lapsing rate to zero or to a small constant; the guessing rate to the expected chance performance—that is, to zero in a yes/no task or to $1/a$, with a being the number of alternatives in a forced-choice task. For the slope parameter, there is often no justification for the particular value used. How well these defaults are met is often not verified. In those cases in which only one parameter (usually the threshold) is estimated, incorrect specification of the other parameters is problematic, since it may introduce an unknown amount of bias to the parameter of primary interest (McKee, Klein, & Teller, 1985; O'Regan & Humbert, 1989). When the function fit is used as the basis for stimulus placement in an adaptive scheme (as,

The study was supported by the Deutsche Forschungsgemeinschaft (Re 337/7) and by the Friedrich Baur Foundation (medical faculty of the Ludwig-Maximilians-Universität München). Correspondence concerning this article should be addressed to B. Treutwein, Institut für Medizinische Psychologie, Ludwig-Maximilians-Universität, Goethestr. 31, D-80336, München, Germany (e-mail: bernhard@imp.med.uni-muenchen.de).

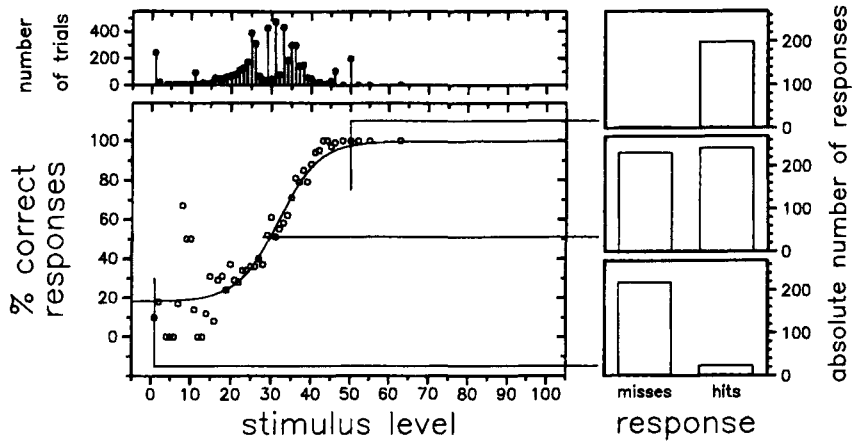

Figure 1. Psychometric function and the underlying binomial distribution at three fixed stimulus values. Top left: histogram of the number of presentations as a function of stimulus level. Bottom left: percentage of correct responses and a fitted logistic psychometric function. Right: The three subplots depict the number of correct/incorrect responses at three specific stimulus values, thereby illustrating the binomial distribution of responses at a fixed level. Data are from the nine-alternative forced-choice task described in the Results section. Here, data have been pooled over 200 different conditions (25 sessions × 8 locations).

for example, in Harvey, 1986, 1997; King-Smith, Grisby, Vingrys, Benes, & Supowit, 1994; Lieberman & Pentland, 1982; Treutwein, 1997; Watson & Pelli, 1983), there is the possibility of adverse effects on the placement strategy (Green, 1990; Madigan & Williams, 1987; Swanson & Birch, 1992).

Although a fit will generally be better the more parameters are left to vary, convergence quickly deteriorates with an increasing number of parameters, and the more so the smaller the data set. Our goal was to develop a routine that provides a stable fit, including those difficult-to-handle cases where all four parameters are determined simultaneously. We were, further, particularly interested in being able to fit data from adaptive procedures in which the available data are scarce—that is, in which only a limited number of trials are performed and in which, as is not the case in a constant stimulus design, each stimulus value occurs only once or a small number of times. We describe the procedure and evaluate its capabilities through Monte Carlo simulations. We also demonstrate its applicability in the reanalysis of data from psychophysical experiments. The method has been implemented as a set of modules[3] that can be called from an application program for postexperimental analysis. With slight modifications and on a modern PC, it may be used online for stimulus placement in experiments with sequential testing.

BACKGROUND

Subject responses in the involved psychophysical tasks usually constitute binary data—the subject's answers can be correct or incorrect or can be yes or no—and the mathematical/statistical problem is, therefore, quite general. Applications and further development of methods are found not only in psychophysics but also in such diverse areas as toxicology, genetics, and materials research, to name a few. There is a wide and profound literature concerning the question of how a cumulative probability distribution can be fitted to binary response data (for textbooks, see Collett, 1991; Cox & Snell, 1989; Finney, 1971; Morgan, 1992). These methods, however, generally do not provide solutions for nonzero guessing and lapsing rates, and they require large data sets.

The history of fitting *psychometric functions* to binary responses can be traced back at least to Fechner (1860/1966), who proposed using linear regression on the response data transformed by the inverse cumulative-normal distribution. Müller (1879) and Urban (1908), as cited by Fechner (1882/1965) and Guilford (1954), modified this approach by introducing weights that are inversely proportional to the expected error of the data. Finney (1971) refined this method by iteratively recalculating the weights to be proportional to the *predicted* variance of the *measured* data. He showed that the parameters resulting from this iteration process converge to the ML parameter estimates when data samples are large. Such methods are now generally subsumed under the topic of generalized linear modeling (GLIM; see Agresti, 1990). In GLIM, the dependent variable is first transformed with the so-called *link function*. In this step, the response data are linearized, and their variance is stabilized. Then, a linear regression on the transformed data is performed. As link function, the methods mentioned so far use the inverse normal distribution function, often called probit transform; the logit and the complementary log–log transform are other common choices. The latter are the inverse to the logistic and the Gumbel (Weibull, on a logarithmic abscissa) distribution, respectively. An empirical evaluation as to which of these functions better fits the data is difficult (see Morgan, 1992, p. 28), and the theoretical implications are debated vigor-

ously. The inclusion of terms for guessing and/or lapsing probability renders the problem intrinsically nonlinear and, therefore, prohibits the use of GLIM methods. Finney himself was aware of this problem and proposed, in the third edition of *Probit Analysis* (1971), a numerical maximization of the log-likelihood.[4]

In sequential estimation of the threshold only, ML methods and Bayesian estimation (based on the mean of the posterior density) are quite common (Emerson, 1986; Harvey, 1986, 1997; King-Smith et al., 1994; Lieberman & Pentland, 1982; Treutwein, 1997; Watson & Pelli, 1983). These methods generally calculate a sampled representation of the likelihood and perform a grid search for its maximum or approximate the integral by a summation of sampled values (Bayes).

To estimate threshold *and* slope, Watson (1979) suggested a two-dimensional ML estimation by a grid search on the two-dimensional sampled representation of the likelihood. Green (1993) used a similar approach for sequential estimation of threshold and *guessing rate* (lower asymptote).

A nonlinear least-squares regression routine, such as the Levenberg–Marquardt method (Brown & Dennis, 1972; Reich, 1992) may be suitable for fitting all the parameters of a psychometric function, but, for that purpose, the method has to be modified to use iteratively reweighted nonlinear least-squares fitting, to account for the fact that the responses are binomially distributed.

To achieve our goal of fitting the nonlinear four-parameter model (threshold and spread, guessing and lapsing rate) with only a small number of trials, we chose a Bayesian approach, since it allows the use of constraints through deliberate specification of a prior distribution for each of the estimated parameters. Bayesian estimation methods are based on the application of Bayes' theorem, which relates the conditional probabilities $P(\cdot\,|\,\cdot)$ and marginal probabilities $P(\cdot)$ associated with two events, say A and B, as follows:

$$P(B\,|\,A) = \frac{P(A\,|\,B)P(B)}{P(A)}. \tag{1}$$

Identifying event A with the data X collected in an experiment and B with the parameter set Θ to be estimated, Equation 1 can be written as

$$P(\Theta\,|\,\mathrm{X}) = \frac{P(\mathrm{X}\,|\,\Theta)P(\Theta)}{P(\mathrm{X})}. \tag{2}$$

Now, $P(\mathrm{X}\,|\,\Theta) = \mathcal{L}(\Theta\,|\,\mathrm{X})$ is the likelihood function (see below), $P(\Theta)$ is the prior distribution specifying our beforehand knowledge about the parameter values—that is, the constraints—and $P(\mathrm{X})$ is a normalization constant for a given experiment.

METHOD

Psychometric Function

A psychometric function $\psi(x;\Theta)$ describes the probability of a correct answer as a function of stimulus intensity x (see Figure 1). Its shape is given through a parameter set $\Theta = \{\theta, \sigma, \gamma, \lambda\}$, where θ denotes the threshold location on the abscissa, σ is a shape parameter describing the spread or inverse steepness of the function, and γ and λ denote the guessing and lapsing probabilities, respectively. Given a sigmoid function $F(x;\,\theta,\,\sigma)$ with asymptotes zero and one—that is, a cumulative distribution—the general psychometric function can be written explicitly as

$$\psi(x;\,\theta) = \gamma + (1 - \gamma - \lambda)\,F(x;\,\theta,\,\sigma). \tag{3}$$

Commonly used cumulative distribution functions are the normal,[5] the logistic, the Gumbel,[6] and the Weibull distributions. In our simulations and reanalysis of psychophysical data, we used the logistic distribution for convenience and without any theoretical implications.

Likelihood of the Data

The likelihood function is defined as the unnormalized conditional joint probability density of the data, interpreted as a function of the parameter set Θ (Collett, 1991, p. 49; Fisher, 1912; Morgan, 1992). Given an experimental session, with n stimulus presentations at intensities $(x_1, \ldots x_n) = \mathrm{X}$, the likelihood function is thus defined as

$$\mathcal{L} = (\Theta\,|\,\mathrm{X}) = \mathcal{L}(\Theta\,|\,\mathrm{x}_1,\ldots,x_n) = \prod_{i=1}^{n} L(\Theta\,|\,x_i). \tag{4}$$

Each $L(\Theta\,|\,x_i)$ in the product is the probability that the subject has given a particular answer—correct or incorrect—when a stimulus with intensity x_i was presented at trial i. This probability is considered for different values of the parameters in the set Θ. The probability of a subject giving a certain answer is obtained from the psychometric function. For a correct answer, this probability is, by definition, given by $\psi(x_i, \Theta)$; for an incorrect answer, it is given by the complement $[1 - \psi(x_i, \Theta)]$. The likelihood functions $L(\Theta\,|\,x_i)$ in the product of Equation 4, each for a single trial i, are therefore given by

$$L(\Theta\,|\,x_i) = \begin{cases} \psi(x_i,\Theta) & \text{if the response is correct} \\ 1-\psi(x_i,\Theta) & \text{if the response is incorrect} \end{cases} \tag{5}$$

For a given data set, there are usually more trials n than different stimulus levels k; let there be l_i trials at level x_i—that is,

$$n = \textstyle\sum_{i=1}^{k} l_i.$$

Let c_i denote the number of correct answers at the stimulus level x_i. The outcome of the complete session will be a specific combination of correct and incorrect answers. The likelihood for the complete data set of a session is the product of the probabilities over all events and is, according to Equations 4 and 5, given by

$$\mathcal{L} = (\Theta\,|\,\mathrm{X}) = \prod_{i=1}^{k} \binom{l_i}{c_i} \psi(x_i,\Theta)^{c_i}\,[1-\psi(x_i,\Theta)]^{l_i-c_i}. \tag{6}$$

The binomial coefficients

$$\binom{l_i}{c_i}$$

take into account that, at each of the different stimulus levels, there are several presentations; the probability at each level is assumed to be independent of the sequence in which the stimuli were presented.

In general, the likelihood will be an extremely small value, since it results from raising a probability (i.e., a value between zero and one) to a high power—namely, the number of trials n. In a purely Bayesian approach, the likelihood is normalized in such a way that the total integral over the likelihood equals *one*, thereby transforming the likelihood into a multivariate probability density for the parameter values. This normalization factor is given by

$$K_B = \int_\Theta L(\Theta) \mid X) d\Theta.$$

The computation of such a four-dimensional integral is tedious, and, since we are only interested in finding the *maximum* of the likelihood function, a precise normalization is not required. We have implemented an approximate normalization, to prevent numerical problems of underflow, that uses a scaling constant of $K = 2^{-n}$ in a yes–no design and $K = (2a/1+a)^{-n}$ in a forced-choice design with a alternatives.

With these ingredients, we can now specify the full relationship for obtaining the posterior probability distribution of the parameter set Θ. Note the identity between the likelihood $L(X \mid \Theta)$ and the conditional probability of the data given the parameters $P(X \mid \Theta)$. Inserting Equation 4 into Equation 2 yields

$$P_{post}(\Theta \mid X) = \frac{1}{K} P_{prior}(\Theta) \prod_{i=1}^{n} L(\Theta) \mid x_i), \tag{7}$$

where K stands for the normalization factor defined above and $P_{prior}(\Theta)$ specifies the prior knowledge about the parameter values. The product of likelihoods depends on the psychometric function, as given in Equation 5.

Prior Distribution

One of the key elements in Bayesian statistics is the use of prior distributions that provide an elegant and explicit way of incorporating beforehand knowledge about a parameter value. Often, it is critically objected that Bayesian statistics thereby introduce an element of subjectivity into a seemingly objective procedure. Although this looks like a valid critique, a closer inspection of the alternative methods shows that they, too, have similar assumptions; they are just less explicit: If we prespecify, for example, a model of a psychometric function where all parameters are fixed except that of threshold, we assume complete knowledge about the three fixed parameters, although that knowledge is not available. Translated into Bayesian terminology, fixing of a parameter corresponds to using a needle-sharp delta function as the prior distribution, thus predetermining the outcome of the estimate. Conversely, those parameters that are left to vary in ML fitting get equal chance for every possible value, which corresponds, in Bayesian terms, to using a rectangular prior. The inadequacy of the latter approach lies in considering all statistical fits as equivalent, including those with nonsensical parameter values, such as a guessing or a lapsing rate above one or below zero. With large data sets, that might not pose a problem, but data points in psychophysical research are commonly scarce—that is, the number of presentations is limited and is far from being sufficient for fitting more than one parameter in that way. Most researchers, therefore, assume complete knowledge about all parameters except that of threshold. The use of a Bayesian prior relaxes the strong assumption of complete knowledge and allows specification of the prior knowledge more appropriately as a smooth function.

What are adequate priors for the psychometric function's parameters? Ultimately, the question of how these psychological variables—threshold, slope, guessing rate, and lapsing rate—are distributed is an empirical one; ideally, the empirical distributions should be determined and used. As a first step, and before full prior data is available, some ad hoc assumptions are in place, though: First, the guessing and lapsing rates need to be kept strictly within the valid range of probabilities—that is, in the open interval (0,1). For the lapsing probability λ, it seems advisable to emphasize small values—say, in the range of 0%–5%. The prior for the guessing probability γ will depend on the experimental design: In a yes–no design, the distribution of γ will be similar to that of the lapsing probability, since, for an ideal observer, γ is zero at small stimulus intensities and may be slightly above zero for some real observer. In a forced-choice experiment with a alternatives, the value $1/a$ is assigned the highest probability. Estimation of guessing rate (rather than fixing it at $1/a$) is meaningful when subjects can have different preferences for the alternatives and averaging over the alternatives is not appropriate for some reason (as, for example, in the experiments shown later).

For the implementation of the priors, an analytical description will be the most compact way; we found the beta distribution $\mathcal{B}(x; p,q)$ well suited. The beta distribution is a probability density function, defined over the interval (0,1). With different values of the parameters p and q, $\mathcal{B}(x; p,q)$ takes a wide variety of shapes, from rectangular over symmetrically or asymmetrically bell-shaped, inverse U-shaped or U-shaped, to L-shaped. Figure 2 shows some examples. According to Martz and Waller (1982), the beta distribution is given by

$$\mathcal{B}(x; p,q) = \mathrm{B}(p,q)^{-1} x^{p-1} (1-x)^{q-1}, \tag{8}$$

where $\mathrm{B}(p,q)$ is the beta function (see Press, Teukolsky, Vetterling, & Flannery, 1992) defined by

$$\mathrm{B}(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1} dt = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} = \mathrm{B}(q,p). \tag{9}$$

The gamma function $\Gamma(x)$, in turn, is defined by the integral

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt. \tag{10}$$

The beta function can be calculated from an approximation of the gamma function given, e.g., in the *Numerical Recipes* (Press et al., 1992, Chap. 6.1).

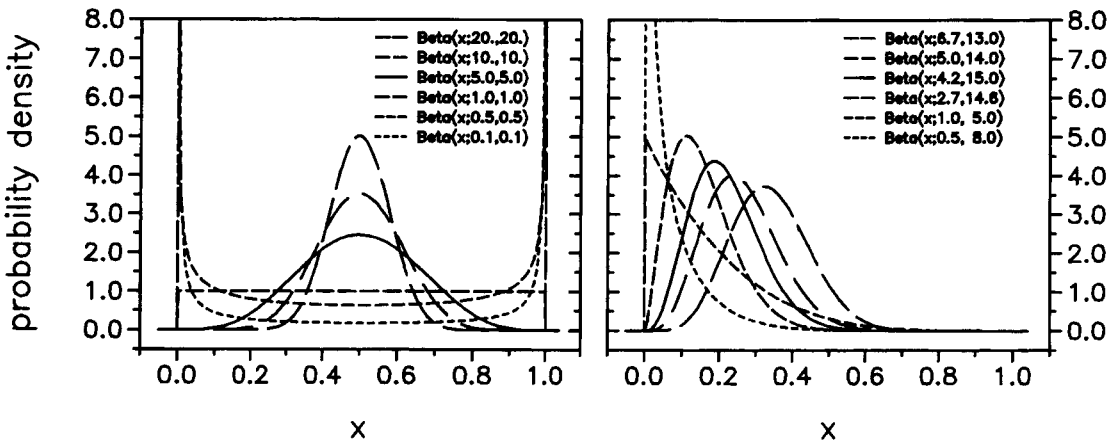

Figure 2. Examples of beta distributions $\mathcal{B}(p,q)$, with symmetrical (left) and asymmetrical (right) shapes.

There is a theoretical justification that lets the beta distribution appear as the prior of choice. According to Martz and Waller (1982), the best suited prior for the estimation of a parameter is the conjugate to the distribution of that parameter. Guessing and lapsing rate can be seen as representing the typical response behavior at minimal and maximal stimulus intensity, respectively. Since responses, at a fixed intensity, are binomially distributed, the conjugate to the binomial is an appropriate prior. The beta distribution *is* that conjugate (see Martz & Waller, 1982, Chap. 6.2) and, on this ground, seems an appropriate choice as a prior for the guessing and lapsing parameters.

Prior distributions for threshold and slope need to exclude impossible values (e.g., negative Michelson contrasts or negative durations), and provide a way to restrict the estimates to those that lie within a reasonable range. How this can be accomplished with beta distributions will be detailed at the end of the section about normalizing the stimulus range.

Maximizing the Likelihood

Finding a function's maximum is equivalent to finding the minimum of the negative of that function; therefore, minimization routines (which are routinely available)[7] can be used for finding the maximum of the posterior distribution. The task is to find the minimum of the negative likelihood[8] $\mathcal{L}'$:

$$\mathcal{L}'(\Theta; \mathrm{X}) = -\mathcal{L}(\Theta; \mathrm{X}). \tag{11}$$

A straightforward minimization method is the multidimensional *downhill simplex method* (Nelder & Mead, 1965; for an evaluation, see Olsson & Nelson, 1975), an implementation of which is readily available in *Numerical Recipes* (Press et al., 1992).The method of Nelder and Mead is one of the rare optimization routines that are claimed to be capable of minimizing nonsmooth functions (see Chandler, 1969a, 1969b). During the development of our method, we started off with nonsmooth constraints, and we kept this routine later for convenience.

Normalizing the Stimulus and Spread Range

Generally, multidimensional routines work best if the data are of comparable magnitude on the different dimensions. Guessing and lapsing probabilities are in the same range (0,1), but threshold and slope, which are derived from physical stimulus properties, depend on the experimental conditions and need to be rescaled to an appropriate range. For the minimization, threshold and slope are, therefore, mapped onto the common interval (0,1). This way of mapping simplifies the specification of priors for threshold and spread (see below). A stimulus range ($x_{\min} \ldots x_{\max}$) that spans the sensory range has to be specified by the user; the mapping from these (physical) values x to internal (optimization) values ξ is straightforwardly given by the linear transform:

$$\xi = A + Bx \quad \text{and} \quad x = \frac{\xi - A}{B}, \tag{12}$$

with A and B obtained by

$$A = \xi_{\min} - Bx_{\min} \quad \text{and} \quad B = \xi_{\max} - \frac{\xi_{\min}}{x_{\max} - x_{\min}}. \tag{12a}$$

The mapped variable ξ might be thought of as an internal excitation level that corresponds to the external physical magnitude x. From Equations 3 and 12, we obtain psychometric functions defined over a thereby normalized range, with $\xi \in (0,1)$ as independent variable. Example plots for the logistic function L,

$$\mathrm{L}(\xi, \theta_\xi, \sigma_\xi) = \frac{1}{1 + \exp\frac{\theta_\xi - \xi}{\sigma_\xi}}, \tag{13}$$

are shown in Figure 3, with both an arbitrary unnormalized and a normalized abscissa. This cumulative distribution is used—in combination with Equation 3—throughout the simulations and also for the fits of the experimental results.

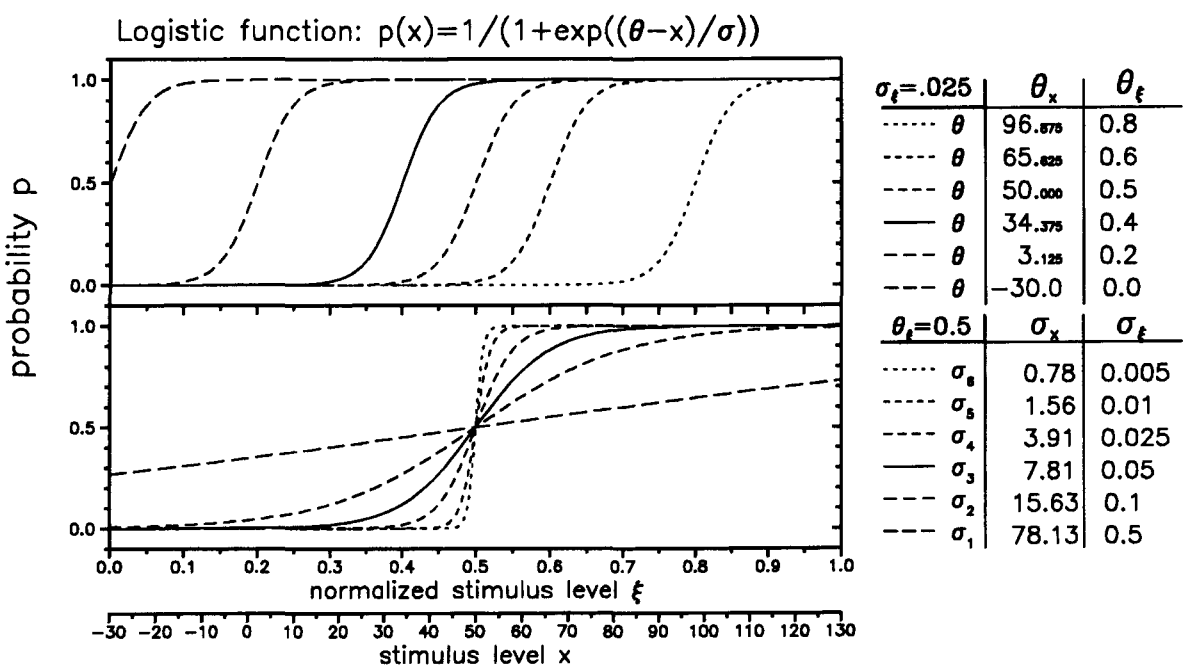

Figure 3. Logistic functions over both an (arbitrary scaled) unnormalized abscissa (x) and a normalized one (ξ), for a number of threshold θ and spread σ (inverse slope) values (the respective other parameter held constant). θ and σ are specified in x-coordinates; see the table on the right, which also gives the corresponding ξ values.

From the lower part of Figure 3, it is apparent that a reasonable range for restricting the spread parameter, in ξ units, is the interval (0.005, 0.5): At a value of $\sigma_\xi = 0.5$, the sigmoid has degenerated to an almost linear function over the considered range, so this upper bound is high enough to include all sigmoids. The set lower bound of $\sigma_\xi = 0.005$ corresponds to an almost step-like function at the chosen resolution after normalization. The even grading of spreads in Figure 3 suggests that the variation of the spread parameter for the maximization is best done in logarithmic scaling. Thus, we transform the spread from the interval (0.005, 0.5) to the interval (0,1), according to

$$\sigma_\xi^{(\mathrm{opt})} = \frac{\log\sigma_\xi - \log 0.005}{\log 0.5 - \log 0.005} = \frac{\log_{10}\left(\frac{\sigma_\xi}{5}\right)+3}{2} = \frac{1}{2}\log_{10}\left(\frac{B\sigma_x}{5}\right)+1.5, \qquad (14)$$

before entering the maximization routine, and transform back, after leaving that routine, by

$$\sigma_x = \frac{0.005\cdot 10^{2\sigma_\xi^{(\mathrm{opt})}}}{B}. \qquad (15)$$

Having transformed the allowable ranges for threshold and spread to values in the interval (0,1), we are now able to specify beta distributions for these two parameters.

Choosing appropriate distribution parameters depends on the experimenter's preferences. Setting p and q to 1 in the beta distribution results in a rectangular, uniform distribution—that is, one that assigns equal prior probability to each of the possible values for the respective psychometric function parameter;[9] this is shown as the short dashed graph in Figure 4. Increasing the values for p,q slightly—for example, to $p = q = 1.2$ (solid curve in Figure 4) or $p = q = 1.4$ (long dashes in Figure 4)—changes the beta distribution's shape from uniform to symmetrically inverse U-shaped with a maximum slightly above 1 in the center region [$\mathcal{B}(0.5; 1.2, 1.2) = 1.1$ and $\mathcal{B}(0.5; 1.4, 1.4) = 1.3$] and a smooth decrease to zero at the borders. The prior probability density for the threshold and spread parameter is then more concentrated in the middle of the definition range. At the values $\xi \approx 0.18$ and $\xi \approx 0.82$, these inverse U-shaped densities intersect with the uniform distribution, and, outside the interval (0.18, 0.82), the densities quickly drop to zero. We propose to use these intersection points as $\xi_{\min} = 0.18$ and $\xi_{\max} = 0.82$ in Equation 12a, thereby extending the allowable range for fitted thresholds to values outside the interval ($x_{\min}, x_{\max}$) specified by the experimenter. This is useful for spotting formal errors: Whenever such values occur, the input response data should be rechecked for consistency.

Final Optimization Problem[10]

From Equations 6 and 7, the complete function that will be maximized is given by

$$\mathcal{F}(\Theta) = \mathcal{F}(\theta,\sigma,\gamma,\lambda) = \mathcal{B}(\Theta)\mathcal{L}(\Theta\,|\,\mathrm{x}) = \frac{1}{K}P(\theta)P(\sigma)P(\gamma)P(\lambda)\prod_{i=1}^{k}\binom{l_i}{c_i}\psi(x_i,\Theta)^{c_i}[1-\psi(x_i,\Theta)]^{l_i-c_i}, \qquad (16)$$

where $P_{\mathrm{prior}}(\Theta) = \mathcal{B}(\Theta) = \mathcal{B}(\theta, \sigma, \gamma, \lambda = \prod_{\zeta\in\{\theta,\sigma,\gamma,\lambda\}}\mathcal{B}(\zeta; p_\zeta, q_\zeta)$ denotes the multivariate prior given as the product of the individual priors for the four parameters $P(\cdot) = \mathcal{B}(\cdot; p., q.)$, as described in the former sections; K is the normalization factor as given above; k is the number of stimulus levels used; l_i is the number of presentations at stimulus level x_i; c_i is the corresponding number of correct responses at the same level; and ψ is the psychometric function as defined in Equation 3. The actual (p,q) values of the beta distributions used in the fitting are shown in Tables 1 and 2; the corresponding beta distributions are illustrated in Figures 2 and 4.

RESULTS

Monte Carlo Simulations

Monte Carlo simulations for two designs of psychophysical experiments will be presented: (1) a single-presentation design—that is, a special version of a constant-stimulus design—with 100 stimulus levels and only *one* presentation at each level; (2) data from a set of simulated adaptive sessions, with stimulus placement according to the algorithm proposed by Kesten (1958), combined with a dynamic termination criterion based on the size of the final Bayesian probability interval (see Treutwein, 1995). For both applications, a psychophysical subject having a psychometric function with a number of different shapes was simulated. The procedure was the following: At some stimulus level x, a single binary response, with a probability for the correct response given by the value of the psychometric function, was generated.[11] This was repeated for different stimulus values. Finally, the fit routine was

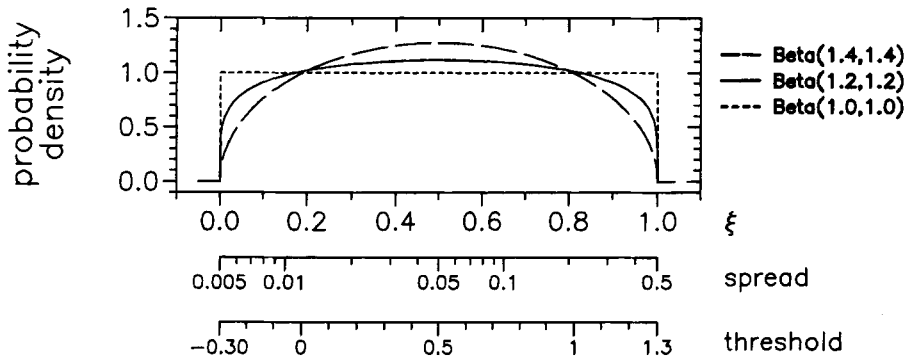

Figure 4. Beta distributions used as priors for the threshold θ and σ spread (inverse slope) parameter, shown as functions over the normalized and original range.

Table 1
Values for the Parameters (p,q) of the Beta Distributions Used in our Fits for Threshold θ, Spread σ, and Lapsing Rate λ

Parameters	θ	σ	λ
(p,q)	(1.2,1.2)	(1.4,1.4)	(0.5,8)

Table 2
Suitable Values for the Parameters (p,q) of the Beta Distribution for the Guessing Rate γ for Different Experimental Designs (Yes/No, Two-, Three-, Four-, Five-, Eight-, and Nine-Alternative Forced Choice)

	γ					
Parameters	Yes/No	2-AFC	3-AFC	4-AFC	5-AFC	8,9-AFC
(p,q)	(1.0,5.0)	(10,10)	(6.7,13)	(5,14)	(4.2,15)	(2.7,14.6)

called with this set of simulated binary responses, to recover the (known) parameters of the simulated psychometric function. By repeating this process, it is, therefore, possible to evaluate the fitting routine with respect to bias and precision in recovering the known properties of the simulated psychophysical subject from limited samples.

Single-presentation design. One of the advantages of our method of fitting binary response data is that it is capable of handling data that contain predominantly single responses—that is, that the responses need not to be pooled before fitting. The single-presentation simulation represents a special variant of the method of constant stimuli, in which each stimulus value is presented only once but a large number of different levels is used. As stimulus levels, we arbitrarily used a range of [1,100]; threshold bias can, therefore, be interpreted as a percentage. We used a $9 \times 3 \times 3 \times 11$ factorial design, with the following conditions: A subject was simulated having its threshold at nine different locations, $\theta \in \{10, 20, 30, 40, 50, 60, 70, 80, 90\}$ and having three possible values of the spread parameter, $\sigma \in \{5, 10, 20\}$; three values for the guessing rate, $\gamma \in \{0, 0.333, 0.5\}$, were used, corresponding to a yes/no, a three-alternative, and a two-alternative forced-choice design; and 11 different lapsing rates, $\lambda \in \{0, \ldots 0.1\}$, in steps of 0.01, were used. These conditions result in 891 different parameter combinations. At each parameter combination, 200 different random samples, or *simulated sessions*, were drawn, each consisting of 100 binomial responses at the different stimulus values $x = [1, \ldots 100]$. The response probability p was given by the corresponding psychometric function. These data were fed into the fitting routine. One prototypical session, with the simulated responses and the corresponding underlying and fitted psychometric function, is shown in Figure 5.

In the next step, these fits to 200 simulated sessions, each consisting of 100 binary responses, yielded a distribution of fitted parameter values. One of these distributions is shown in Figure 6, for the same underlying psychometric function as that in Figure 5. There is a certain variability of parameter estimates, but the four biases of the parameters are all small. For the different combinations of true parameter values, the simulations yielded 891 distributions similar to those in Figure 6. To assess the fitting routine's precision, two measures from each of these distributions are useful: the deviation of the estimate's mean from the prescribed (true) value (i.e., the bias B) and the

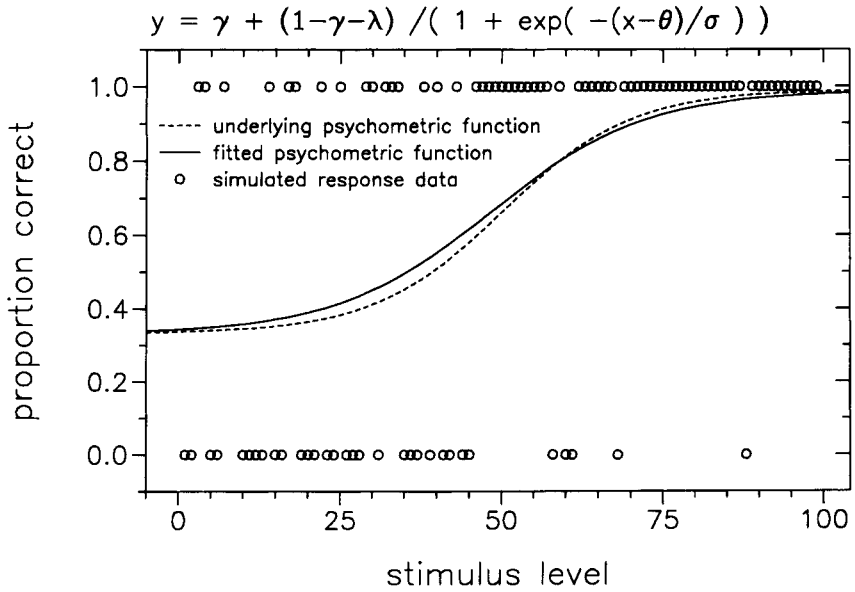

Figure 5. One session of a single-presentation design (simulated response data); proportion of correct responses versus stimulus level. Dashed curve: underlying psychometric function with parameter set $\theta = 50$, $\sigma = 10$, $\gamma = 0.333$, and $\lambda = 0.01$. Open symbols: simulated responses, determined in their proportion of correct/incorrect by the former. Solid curve: fitted psychometric function with parameter set $\theta = 48.1$, $\sigma = 11.9$, $\gamma = 0.332$, and $\lambda = 0.012$.

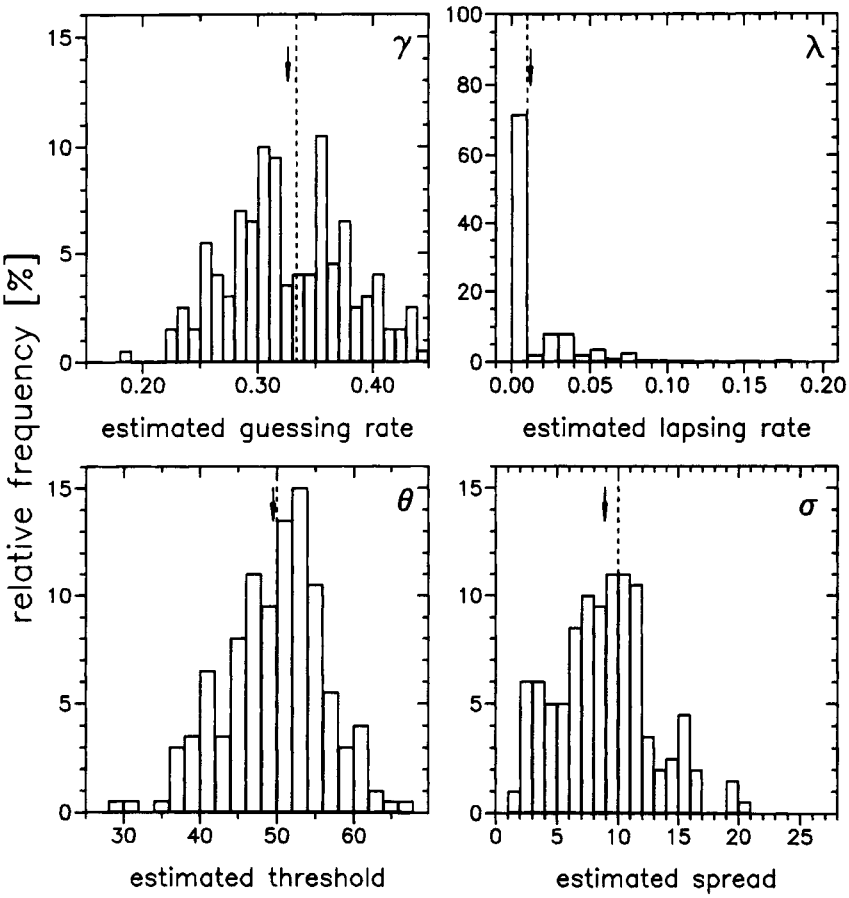

Figure 6. Distribution of the best-fitting parameters in 200 simulated sessions, with 100 single presentations at different stimulus values. The simulated subject had a psychometric function with the parameter values $\theta = 50$, $\sigma = 10$, $\gamma = 0.33$, and $\lambda = 0.01$, indicated by the dashed lines in each subgraph. The mean of each distribution is indicated by an arrow. The estimate's bias is seen as the distance between the arrow and the dashed line; its variability is reflected in the width of the distribution.

estimate's standard deviation S. The bias shows the systematic deviation between the fitted parameter and its true value; it is seen in Figure 6 as the distance between the arrow and the dashed line. The estimate's standard deviation is a measure of the random error introduced by the estimation process. It is seen in the figure as the width of the distribution. By subtracting the true values from the estimates, we get distributions of estimate deviations from the true value (not shown), which have the same shape but are shifted along the parameter axis such that the true values are at zero. The mean and the standard deviation of that distribution of *estimate deviations* are the two measures of interest—bias B, which is a measure for systematic misestimation, and variability S, which reflects the random estimation error, respectively. Note that the standard deviation of estimate *deviations* is equal to the standard deviation of the estimate itself; these two terms are, therefore, used synonymously. The distributions of these two measures are shown in Figures 7 and 8, respectively: Figure 7 shows the distribution of estimate biases B_p for the four parameters $p = \theta, \sigma, \gamma$, and λ, and Figure 8 shows the distributions of those estimates' standard deviations, S_p. Note that these are second-level statistics: Each bias or standard deviation in Figure 7 or Figure 8 is itself based on a distribution like that in Figure 6. Taking all 891 conditions together, a total of around *18 million responses* were simulated in this part of the evaluation.

Concerning the accuracy of the estimates, the best results are obtained for the threshold and spread parameter. A majority of threshold biases lies between ±2% of the range [1..100], and all threshold biases are between −6% and +8%. Spread estimate biases are all between −4 and +5, with a majority being between ±2 (for an interpretation of these values, see Figure 3, bottom, the graphs for σ_3, σ_4, and σ_5). Guessing rate estimate biases come out equally well. Perhaps less convincing are the

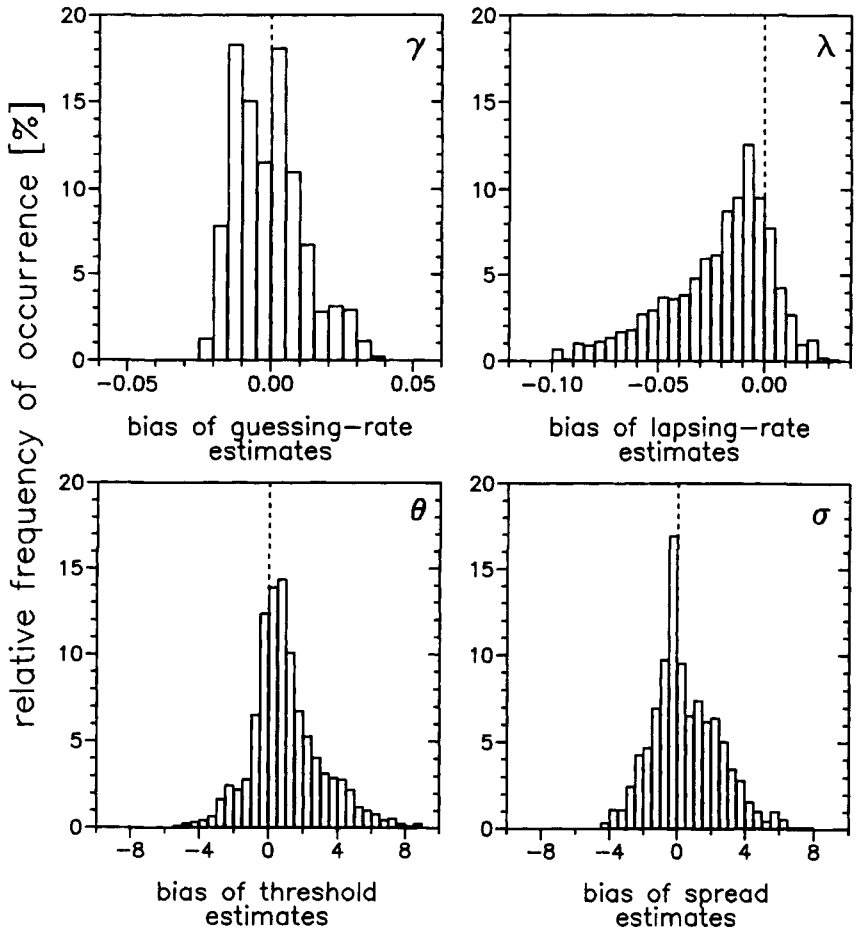

Figure 7. Distribution of biases B_p (i.e., of the systematic deviation of the estimated from the true parameter value) for all parameters p: guessing rate γ (upper left); lapsing rate λ (upper right); threshold θ (lower left); and spread σ, (lower right). The histograms are based on just under 900 conditions of 200 simulated sessions each (and thus based on 18 million simulated responses), so, in the figure, a relative occurrence frequency of 10%, for example, corresponds to an average of around 90 conditions, or 18,000 estimation biases.

results for the lapsing rate biases: Given that the lapsing rate will typically be less than 5%, the overall range of 13% appears quite large.

Although there is little systematic error (as can be seen from Figure 7), Figure 8 shows that the results for random error are less favorable: Although the variation of threshold estimates, for example, can be as low as 2% of the range [1 ... 100], higher values are frequent, up to 14% and more. Similarly, spread estimation can have standard deviations of up to 14–16, although there are also those lucky cases in which the standard deviation is below 2. Guessing rate and lapsing rate estimation is more successful, with a majority of standard deviations lying around 5%. A reason for these results is readily apparent: Stimuli, in this design, are evenly spaced along the stimulus axis. On the one hand, comparatively few stimuli are located around the threshold, where they contribute to the estimation of threshold and spread. A majority of stimuli, on the other hand, are located away from the threshold—that is, in a region where they contribute to the estimation of the guessing and lapsing rates.

The bias distribution for the lapsing rate (Figure 7, upper right) and the distribution of standard deviations for the other three parameters (Figure 8, threshold, spread, and guessing rate) are pronouncedly nonnormal. The question arises, what this implies and where this nonnormality stems from. There might be a systematic dependency on somehow unfavorable parameter constellations for which the estimates for different parameters covary with each other. Another explanation might be influences of the range boundaries: When the simulated data set does not cover the relevant range of the psychometric function, only part of the ogive is matched by stimuli—for example, when Θ is 10 or 90. To detect such influences of parameter constellations, the data shown in Figures 7 and 8 were regrouped, as is illustrated in Figure 9 and de-

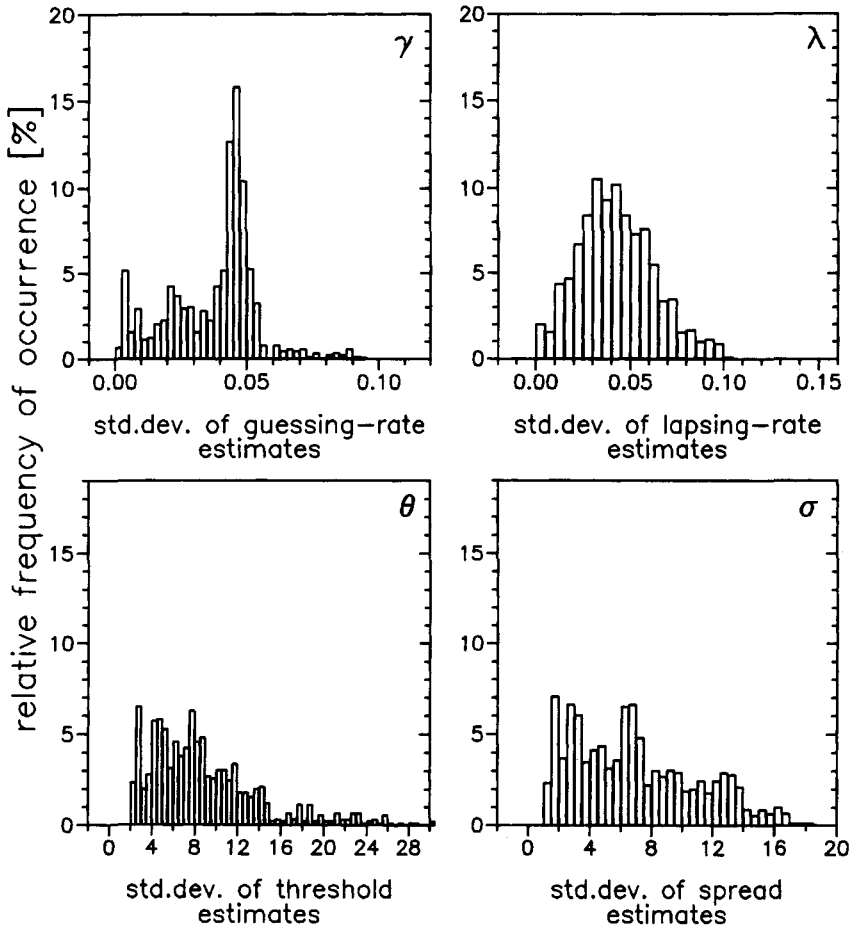

Figure 8. Distribution of the estimate's standard deviations S_p—that is, of the measure for the random variability of the deviations between estimated and true parameters. Layout of subfigures is the same as that in Figure 7.

scribed below, and were subsequently replotted in Figures 10 and 11. Each subplot in the latter two figures represents the variation of a simulated condition and the influence of this variation on the fitted parameters.

The grouping rule by which these distributions were formed was to collect all points that share the value of the underlying parameter that is varied (i.e., that on the abscissa) and to thereby collapse over all remaining parameter combinations. Each subplot in Figures 10 and 11 thus contains *all* datapoints, in a different arrangement each. To illustrate the regrouping, Figure 9 shows threshold distributions grouped by the underlying threshold value as histograms, together with the total histogram in the figure's top. Above each histogram in Figure 9 is a so-called median-box plot that summarizes the histogram: The center is the median of the data falling in this group; the upper and lower box edge delimits the upper and lower hinge, respectively, a hinge being the median of the (upper or lower) half of all points; the "whiskers" delimit the full range of the values. The median box is a graphical representation of Tukey's (1977) *five number summary* (minimum, first quartile, median, third quartile, maximum). In Figures 10 and 11, these median boxes are overlaid by the data points contained in that group (the latter overlapping considerably). These two figures illustrate the errors introduced by the sampling (only a finite number of responses was simulated) and the fitting process: Figure 10 shows the bias (i.e., the systematic error), and Figure 11 the standard deviation (i.e., the random variability).

In Figure 10, consider first those subgraphs that lie on the main diagonal and show the dependency of a parameter's estimation bias on the true value of that same parameter. Although there is little effect to be seen for the parameters threshold, spread, and guessing rate, there are drastic effects for the lapsing rate (top right). Lapsing rate bias shows a steep, approximately linear variation with the true value, where more negative biases are coupled to the higher true values. Higher lapsing rates—under the given circumstances—are thus harder to estimate, with the estimated psychometric function's high asymptote being systematically higher than the prescribed value

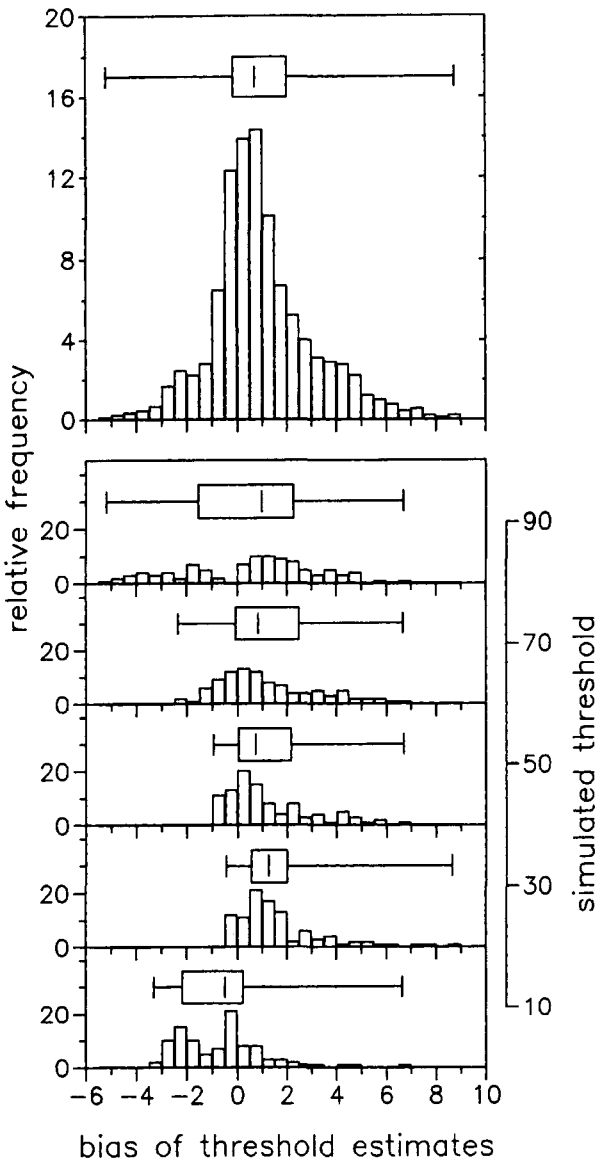

Figure 9. Illustration of the regrouping used to uncover covariation of parameters. Top: distribution and median-box summary of threshold biases, collapsed over all simulated parameter combinations. The histogram is a replot of the lower left subplot in Figure 7. Bottom: distribution and median-box summary of threshold biases for five selected simulated thresholds ($\theta \in \{10,30,50,70,90\}$, see right axis). The median box is a graphical representation of Tukey's *five number summary*; see the text for details.

(λ being systematically underestimated). At the same time, biases are not only larger but are also increasingly variable the higher the true value is, and, for true lapsing rates of 5% and above, there is not a single case of zero bias (i.e., of correct estimation).

The reason lapsing rates are underestimated is simple: The prior was chosen as being plausible for an attending observer, who rarely lapse. The tested range of true values for the lapsing rate extended up to 10%, that limit reflecting an inattentive observer, to see what happens in those unexpected cases. High lapsing rates are preserved by the fit, although they come out smaller than the corresponding true values.

The conclusion to be drawn from the graph is that the overall negative bias seen in Figure 7 for the lapsing rate estimates largely stems from those cases in which the true lapsing rate values were untypically large. Our fitting routine systematically underestimates large true lapsing rates. This can be inferred from comparing the top right graph of Figure 10 with the others in the same row: None of these other three graphs for the lapsing rate show any systematic pronounced variation with the independent parameter, except for some effects from thresholds between 70% and 90% of the scale, discussed below. There is, of course, in all those three graphs an overall negative bias apparent, but that is just a reflection of the fact that each graph contains the full data set—that is, it just represents a regrouping of the data.

Consider next the rightmost column of graphs in Figure 10, which show how true lapsing rate influences the estimates of the other parameters. Again, the influence is pronounced at too high true values. There is little influence on guessing rate estimation, which is not surprising, since guessing rate is reflected in different responses than is lapsing rate. Spread and threshold bias, however, clearly show a systematic dependency on true lapsing rate, with larger spread and threshold biases for the higher (unfavorable) true lapsing rates.

For the illustration of these results, a typical situation of misfit is shown in Figure 12. True threshold, spread, and guessing rate parameters have been chosen as the middle ones of the tested ranges and are representative of these simulations; lapsing rate has been chosen at the unfavorable extreme of 10% (solid lines). A typical fit is shown (dashed lines), with fitted parameters chosen, from Figure 10's rightmost column, as the medians at $\lambda = 10\%$.

The deviation between true and fitted psychometric functions in Figure 12 is small. But since the deviations in these simulations are systematic, they are of interest. The result that a reduced slope and increased threshold accompany the underestimated lapsing rate can be intuitively understood as follows: The inadequate prior, effectively, increases the overall span ($1 - \gamma - \lambda$) of the psychometric function by approximately 5%. Response behavior is inherently ambiguous, in that wrong responses can stem both from incomplete perception (shallow slope) and from insufficient concentration on the task (lapsing). Through the underemphasis of lapsing by the prior, wrong responses tend to be attributed by the ML algorithm to incomplete perception rather than to lapsing. This will be most apparent above threshold, where more correct than incorrect responses are present and the wrong attribution thus flattens the psychometric function and increases its spread and threshold. Put formally, spread and threshold

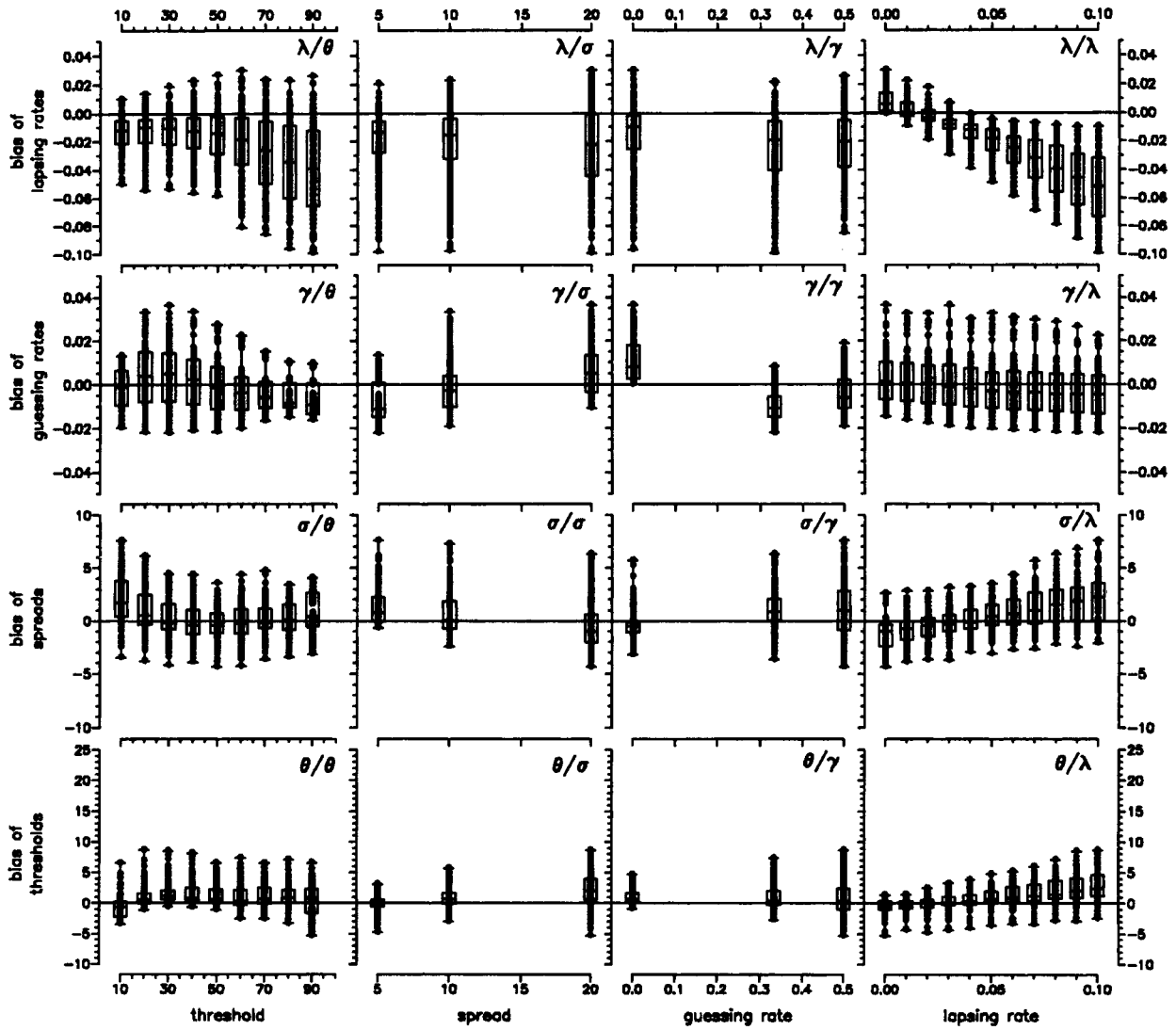

Figure 10. Replot of the data of Figure 7, to show how the *estimation bias* depends on the true parameter values. Each row of subgraphs shows the bias for a certain parameter; each column of graphs shows the dependency on the true value of a certain parameter. Note the markings in the upper right corner of each subgraph. Each graph shows the complete data set, regrouped to show the dependency on some parameter, with all cases that share a true value of a certain parameter (but can arise from different values of the remaining three parameters) arranged in a vertical (bar-like) distribution. See Figure 9 and the text for an explanation of the regrouping and the bar-like representation.

covary with lapsing rate in the optimization space such that a bias in one parameter leads to an opposite bias in the covarying parameters.

Returning to Figure 10, a final effect that needs mentioning is the reduced accuracy of lapsing rate estimation when thresholds are high, between 70% and 90% (top left graph). This can be readily understood by realizing that a high threshold implies a situation in which the stimulus range is insufficient to let the psychometric function go "all the way up," where lapsing would be apparent.

Figure 11 shows the *random* errors of estimation, in a similar arrangement as that in Figure 10 that shows the *systematic* errors. The figure in many aspects confirms the conclusions drawn from the previous figure. Threshold estimation (bottom left graph) works well over a wide middle range of true values, with increased variability at the high extreme of true threshold at 90, although the bias there is still low. Smaller spreads (steep slopes) are more accurately estimated than are larger ones (graph σ/σ). It is interesting that a guessing rate of zero (graph γ/γ) is overall better estimated (small median of variability) than are larger guessing rates but that there is, at the same time, a large number of outliers with high variability at a guessing rate of zero. The lapsing rate (top right graph) shows the corresponding influence on variability to that on bias, high lapsing rates leading to increased variability of esti-

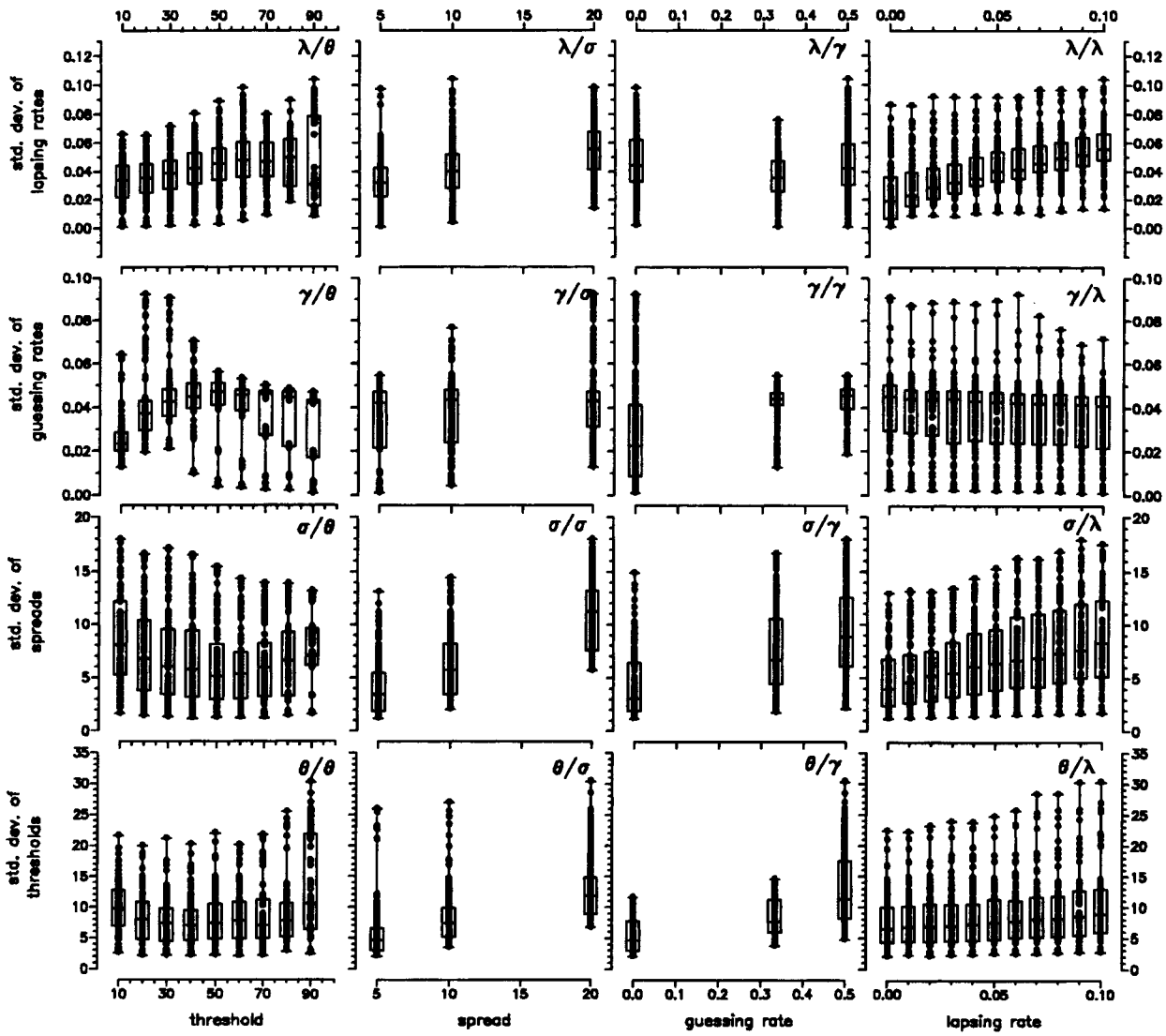

Figure 11. Regrouping of the data from Figure 8, similar to Figure 10. It shows how the random variability of the estimates depends on the true parameter values.

mation. Too high a lapsing rate further decreases the accuracy of spread and threshold estimation (right column of graphs). Finally, note the increased number of outliers in guessing rate estimation when the threshold is low (graph γ/Θ)—that is, when there are few responses below threshold on which to base estimation.

Figures 11 and 10 can be summarized in the following general hypotheses: (1) the range of stimulus values used in the experiment should cover the range of major variability of the psychometric function; (2) the more the prior distributions misrepresent the distribution of the true parameter values, the more bias is introduced in the fitting of that parameter (our simulations show this for the case of the lapsing rate parameter); (3) a failure to fit one parameter is accompanied by failures to fit other parameters that covary in the optimization space; and (4) spread, threshold, and lapsing rate estimates covary with each other.

Adaptive stimulus placement. In a second set of Monte Carlo simulations, we have studied the properties of our fitting technique with data from an adaptive stimulus placement strategy. From the standpoint of psychometric function fitting, such data pose the problem of being unevenly distributed along the stimulus axis—a natural consequence of the optimization toward efficient determination of thresholds. As adaptive placement strategy, we have chosen accelerated stochastic approximation (Kesten, 1958), which, although rarely used in psychophysics, holds promise of being a particularly efficient variant, thus posing a critical test for our fitting routine. Since, in Kesten's original publication, it was left unspecified how the threshold estimate should be obtained (other

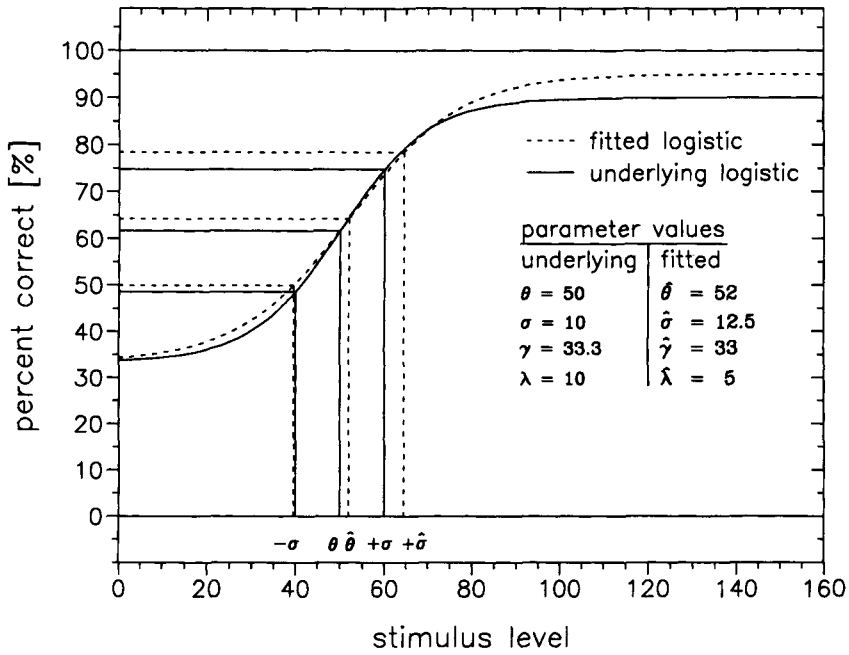

Figure 12. A typical situation with pronounced misfit. True threshold, spread, and guessing rate are chosen to be the middle ones of those tested, but lapsing rate λ is chosen to lie at the extreme. The underlying function and its parameters are shown as solid lines. The fitted function in the particular case shown (dashed line; parameter values indicated by dotted lines) has too low a lapsing rate, too high a spread (i.e., too shallow a slope), and too high a threshold.

than just taking the last stimulus value in the series, which is a rather rough estimate when runs are short), we used our method to calculate final estimates for all four parameters of the psychometric function. Kesten's placement rules are a special variant of a staircase, or up–down, procedure: The session is started with a reasonable initial step size c at some plausible stimulus intensity x_1. For the second presentation, the stimulus intensity is decreased by $c(1 - \phi)$ if the response was correct ($z_1 = 1$) and is increased by $c\phi$ if the response was incorrect ($z_1 = 0$). Parameter ϕ stands for the desired level of performance at threshold (for example 0.5 for a yes/no design or 0.75 for a two-alternative forced-choice design). The rule implies different step sizes for upward and downward steps, except for the special case of $\phi = 0.5$ (i.e., for a yes/no design with zero guessing and lapsing rates). After the second presentation, the step size is diminished whenever a shift in response category occurs (from correct to incorrect or vice versa). The placement of the stimulus levels can be described by

$$\mathbf{X}_{n+1} = \mathbf{X}_n - \frac{c}{1 + m_{\text{shift}}}(\mathbf{Z}_n - \phi), \qquad (17)$$

where, $\mathbf{X}_n$ is the stimulus level presented on trial n, $\mathbf{Z}_n$ is the subject's response on that trial (i.e., 1 for a correct and 0 for an incorrect answer), c is an initial step size, and m_{shift} is the number of shifts in response category. This sequence is guaranteed to converge to x_ϕ—the stimulus level x at which performance level is ϕ—if the number of presentations n is large enough (see Kesten, 1958; Treutwein, 1995).

Distribution histograms of the four estimated parameters, for simulations of a two-alternative forced-choice and of a yes/no design, are presented in Figures 13 and 14, respectively. The simulated sessions were terminated when a certain level of confidence for the threshold parameter was reached (see Treutwein, 1995, p. 2517; Treutwein, 1997, Equation 4). This confidence level was deliberately set low to see how useful results with a rather limited number of trials are. With the set termination criterion, the sessions terminated, on average, after 18.2 trials for the yes/no design and after 63.1 trials for the two-alternative forced-choice design. In both designs, the threshold, guessing rate, and lapsing rate all come out accurately—that is, without a systematic deviation. The variability of threshold estimates is probably a little high for some applications, and one would use more trials when this is the case. The results show, however, that even as few as 20 responses per session allow a reasonable estimation of threshold and simultaneous estimation of further parameters. The spread parameter is less well estimated,

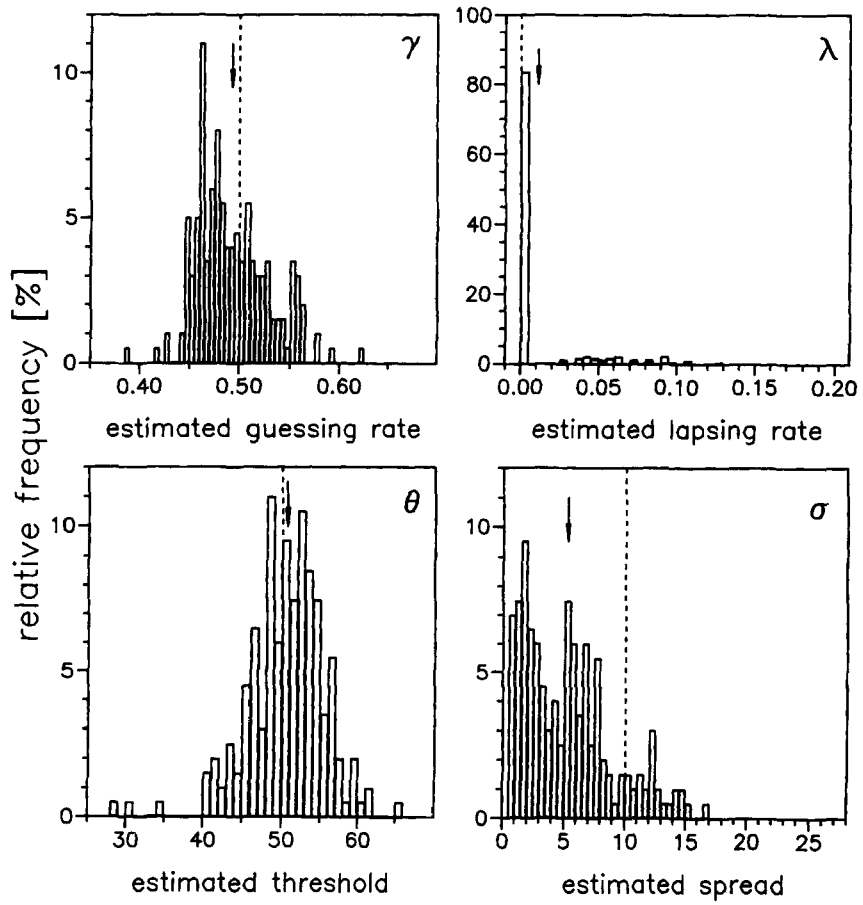

Figure 13. Simulated sessions from a procedure with adaptive stimulus placement (accelerated stochastic approximation) in a two-alternative forced-choice design. Frequency distributions of the fitted parameter values for 200 repeated sessions. The underlying parameter values are indicated by dashed vertical lines; the distribution means are indicated by arrows. (Note the different scaling of the lapsing rate ordinate.)

however, the estimate showing a prominent negative bias in both designs. We have no explanation for why steep rather than shallow psychometric functions are predominant. A large variability of the slope or spread is to be expected, though, given the placement strategy, which is optimized for the estimation of threshold: The adaptive strategy tries to place the trials at the location of the threshold—that is, at a single point—so that slope estimation is unreliable (if the strategy were completely successful to present all stimuli there, the slope would be undefined, and *any* value of the slope parameter would satisfy the fit). If the goal is to determine slope more precisely and still efficiently, a different adaptive strategy is required, preferably one with a bimodal placement distribution centered around the expected threshold (see King-Smith & Rose, 1997).

Application to Experimental Data

Monte Carlo simulations provide a systematic and unambiguous way of evaluating the proper working of a statistical procedure. Nevertheless, data from real experiments often pose unexpected difficulties for any new statistical method, and, ultimately, a method's worth will be demonstrated by wide application. As a first step, the new fitting procedure has been applied for post hoc analysis of data from several psychophysical tasks, one of which will be presented here. The data stem from a spatial nine-alternative forced-choice task for obtaining a certain measure of temporal resolution of human visual performance, the so-called double-pulse resolution. The task was to detect a temporal flicker in one of nine light spots presented at different locations in the visual field. Eight of the locations were arranged on the circumference of a circle, and one in its center; the subjects had to identify the location of the randomly chosen flickering target stimulus. The method and results are described in Treutwein and Rentschler (1992). Here it suffices to state the expected range for thresholds, which is from 0 to 100 msec, and to note that data should not be pooled over the nine alternatives for two reasons: For one, double-pulse resolution (i.e., threshold) is variable across the visual field, and second, a possible response bias toward one of the

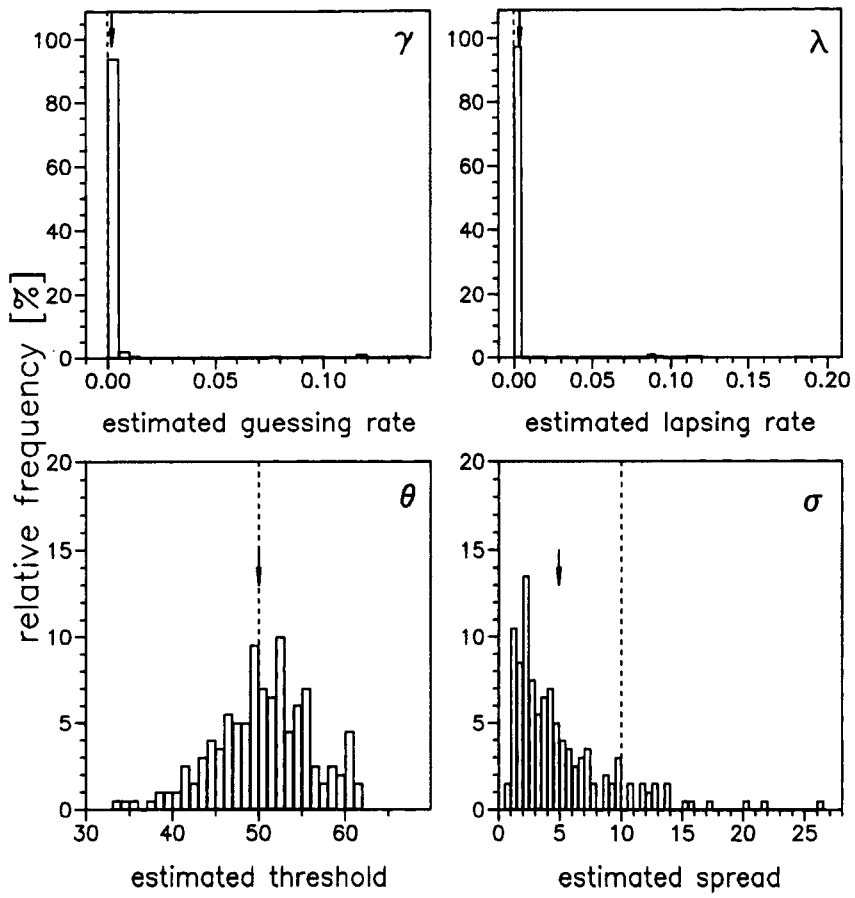

Figure 14. Same as Figure 13, but for a yes/no experimental design.

nine different alternatives would lead to a relatively increased guessing rate for that alternative. Although the subjects were instructed to use *all* the alternatives when guessing, each subject *did* exhibit such bias; the preferred location for guessing was different between subjects.[12] The stimulus levels—in this task, the duration of the target's dark interval—were chosen by an adaptive procedure of the Bayesian type (YAAP; Treutwein, 1997). On average, 36 presentations (33–42) were necessary at each stimulus location in a session to reach the threshold's set reliability level. Sessions were repeated many times under identical stimulus conditions for the same subject to obtain information about the variability of the psychometric function; the data from 25 repeated sessions are shown in Figure 15. Several sources contribute to the overall variability: (1) variability (random error) of estimation, which is of primary interest here; (2) variability of the subject's performance, owing to fatigue, for example; and (3) systematic changes in the subject's performance—for example, through learning.

Most resulting psychometric functions in Figure 15 resemble each other quite well. There are, however, also some remarkable outliers: In some cases, the slope is markedly reduced; in others, the lapsing rate is increased. Since, for real psychophysical data, we cannot know the underlying data-generating process and since it is difficult to define goodness-of-fit measures for sparse and unbinned binary data (see Agresti, 1990, chap. 4.6), we give a graphical representation of the goodness of fit in Figure 16. For this purpose, we took, as an example, the 25 data sets of the lower left panel of Figure 15—which contains one case of reduced slope and one of increased lapsing rate—and plotted, separately for each session, the fit and the corresponding raw data. Each subplot in Figure 16 shows the cumulated rate of correct answers in its lower part and the number of presentations in the upper part; the case of increased lapsing rate is the one on bottom left, and that of reduced slope is the next to the right.

One can see that the increased lapsing rate in that estimate results from one single wrong response (Figure 16, bottom, left graph). The one case of reduced slope in the example (Figure 16, bottom row, second column) is accompanied by stimulus levels spreading over an untypically large range. Some other reason outside the perceptual processes (perhaps a failure of the equipment, an inadvertent change of viewing distance, or sleepiness of

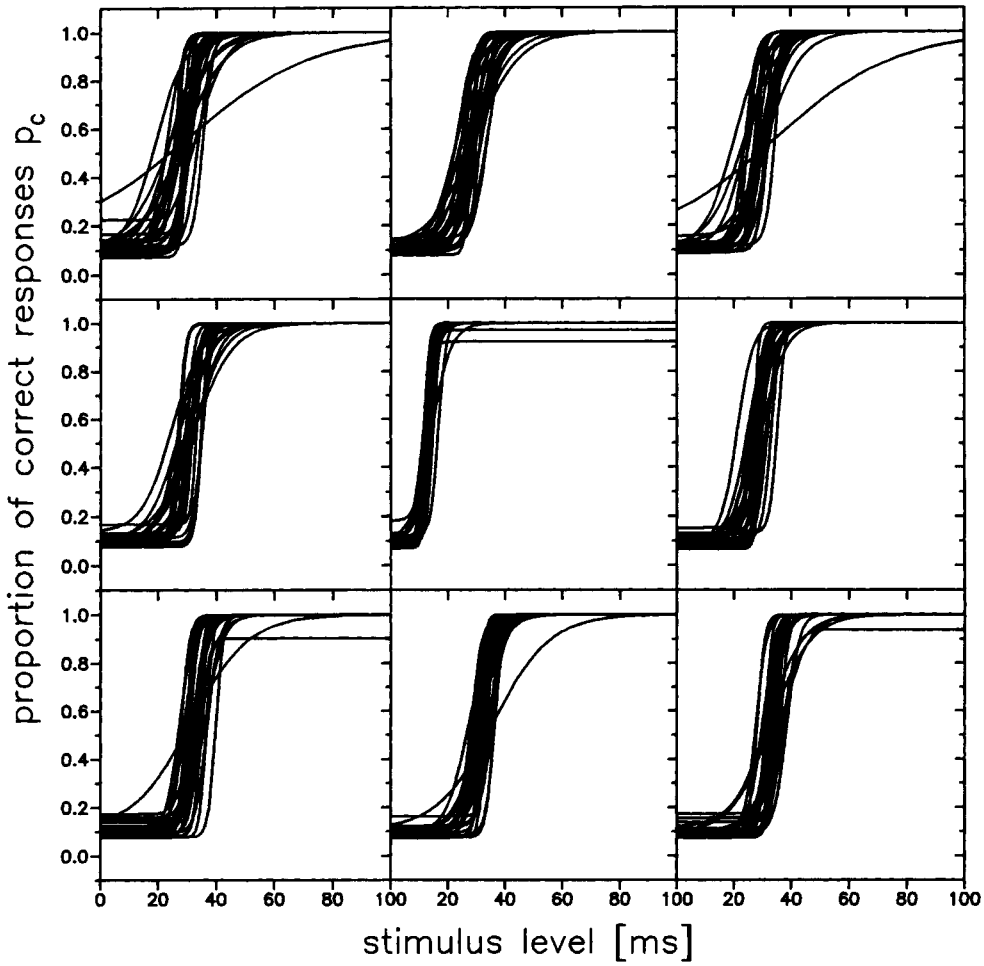

Figure 15. Fitted psychometric functions to repeated measurements in a psychophysical nine (spatial) alternatives forced-choice task, described in the text. The subgraphs are spatially arranged analogous to the stimulus positions in the visual field—that is, the outer eight graphs correspond to the eight peripheral positions on a circle around the middle position and the center graph refers to foveal presentation. Each subplot shows 25 logistic psychometric functions, each the best fitting to the data from a single session. For the nine locations together, these functions represent 225 fits.

the subject) may have contributed to incoherent response behavior; the data set might be given further attention or ultimately excluded from further analysis.

DISCUSSION

With Monte Carlo simulations, we were able to show that as few as 20 trials are sufficient to recover the parameters of a simulated subject in a yes/no task. For scarce data collected with an adaptive placement strategy, considerable bias for the spread is to be expected. Whether this failure to precisely recover the spread parameter is a consequence of assessing all four parameters simultaneously, is a consequence of the adaptive strategy, or is inherent in the ML method, is difficult to decide. There is some evidence in the literature (McKee et al., 1985; O'Regan & Humbert, 1989; Swanson & Birch, 1992) that favors one of the latter two views, but there is, as yet, not enough evaluation of the ML method published for psychophysical application. Available studies fail to cover a comparable range of parameter values and fail to assess systematic interdependencies between threshold and slope, as well as the implication of mismatched guessing and lapsing rates. The data by Swanson and Birch show that erroneous specification of the lapsing rate introduces large bias and a large standard deviation for the estimated threshold values. A possible cause for this bias and variability of the estimated threshold is that the lapsing rate is fixed at a wrong value. Simultaneous estimation of the lapsing rate with our routine will reduce the bias.

As in all multidimensional fits, the parameter estimates covary with each other in the optimization space. For the present case, this means that all four parameters of a psychometric function can only be estimated with

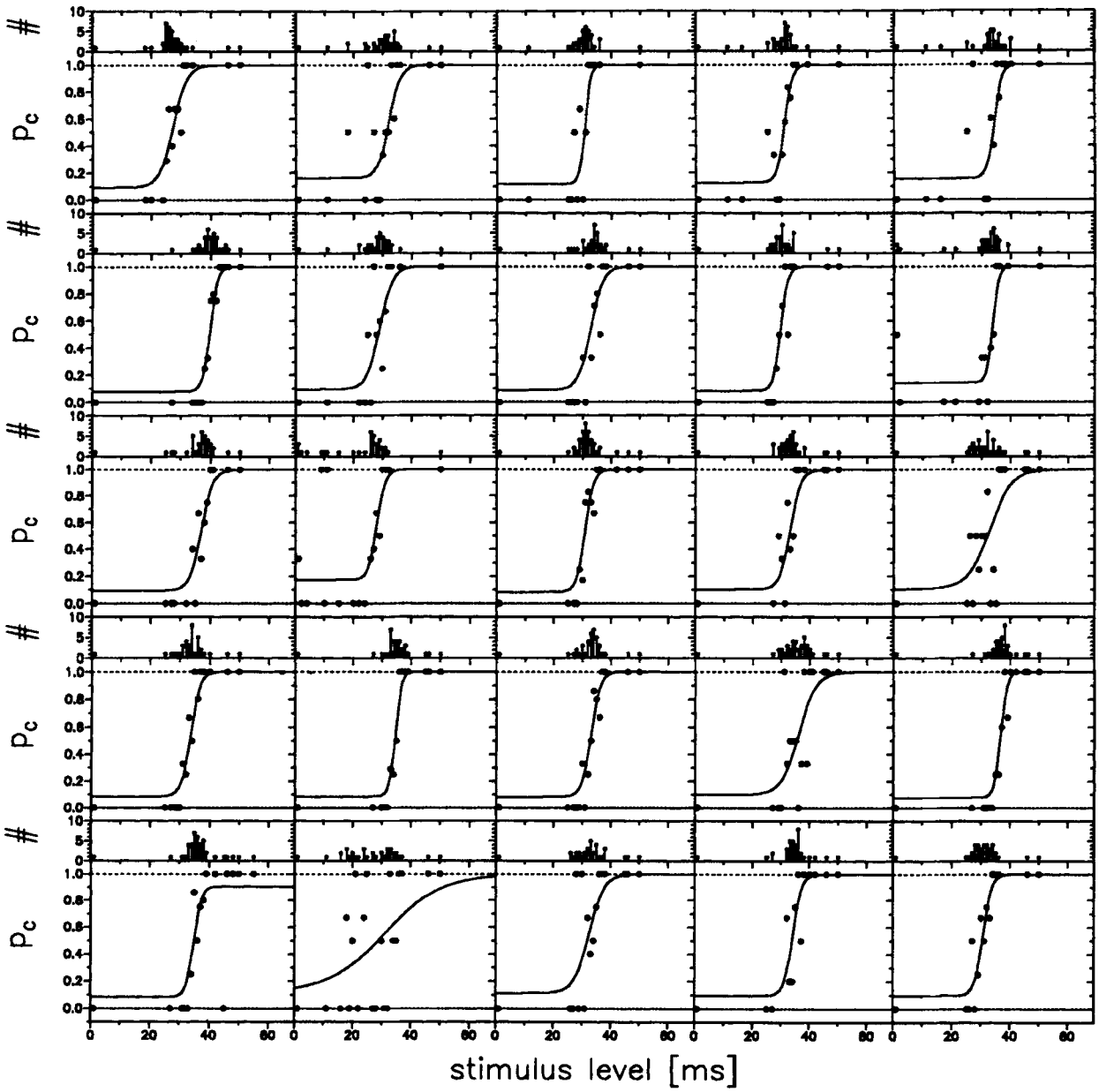

Figure 16. Repeated psychophysical measurements: Data from the bottom left graph in Figure 15 (i.e., 25 repeated sessions at fixed stimulus conditions). Each of the 25 subfigures shows the number of presentations (top), the raw data cumulated at each stimulus level (closed circles in bottom part), and a fitted psychometric function (solid line), all as a function of stimulus level.

some interdependence between the estimates. Although this might be undesirable in very special cases, the advantage of our method is that the rigid assumptions of complete knowledge of the other parameters in the conventional threshold-only fits are relaxed. Our results show that simultaneous estimation of the lapsing rate improves the accuracy for threshold and slope estimation when an appropriate prior for the lapsing rate is chosen.

Although the Bayesian approach is often debated religiously, we think that its application provides a straightforward and reasonable way to realize constraints for a problem that is otherwise difficult to solve. We have chosen beta distributions for specifying the prior information. Another choice is that of penalized likelihoods, but this approach is counterintuitive in that it weights high and low guessing and lapsing rates as equally probable. We find the deliberate construction of prior distributions that reflect the beforehand knowledge the more appealing. The selection of the family of beta distributions as priors for the guessing and lapsing rates is reasonable on the

basis that the beta distributions are the conjugates to the binomial distribution; psychophysical responses at a fixed stimulus value are binomially distributed. For threshold and spread, we lack a theoretical foundation for the specific choice of the beta distribution and have chosen it for convenience. The priors for the latter two parameters should ideally be specified with empirically determined values of the beta distribution's (p,q) parameters such that the latter reflect the prior knowledge—that is, the distribution of threshold and spread estimates in a representative group of the population. King-Smith et al. (1994) used priors in their adaptive placement strategy for threshold measurement (ZEST) that are based on analytical approximations to histograms (19.000 and 70.000 threshold, for two different experimental tasks).

Conclusions

We have developed a constrained generalized ML (CGML) method to fit all four parameters of a descriptive psychometric function—namely, threshold, spread, guessing rate, and lapsing rate. Given the binary nature of the responses, the existence of nonzero asymptotes (guessing and lapsing rates) creates certain difficulties for parameter estimation. We give an overview of the theory behind the method, which is scattered over different areas in the statistical literature. Our method is a constrained extension of a later variant of Finney's (1971) probit analysis, which maximizes the log-likelihood numerically. The method is capable of analyzing binary-response raw data and can be used with sparse data (20–100 trials, depending on the experimental design, seem to be sufficient). Possible applications are the final parameter estimation in experiments using adaptive strategies for the stimulus placement like standard staircase procedures—where often the only estimate is that of averaged reversal points (see, e.g., García-Pérez, 1998; Kaernbach, 1991; Wetherill, 1963) or stochastic approximation where the estimate is the last tested value (Kesten, 1958; Robbins & Monro, 1951). To demonstrate our fitting method's dependability and applicability, we performed *extensive* and *careful* simulations with a broad range of simulated parameter values. The method was further tested with data from psychophysical experiments. Our results show that it is possible to fit all four parameters of a psychometric function dependably.

REFERENCES

AGRESTI, A. (1990). *Categorical data analysis*. New York: Wiley.

BROWN, K. M., & DENNIS, J. E. J. (1972). Derivative free analogues of the Levenberg–Marquardt and Gauss algorithms for nonlinear least squares approximation. *Numerische Mathematik*, **18**, 289-297.

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.

CHANDLER, J. P. (1969a). SIMPLEX—finds local minima of a function of several parameters. *Behavioral Sciences*, **14**, 82.

CHANDLER, J. P. (1969b). STEPIT—finds local minima of a smooth function of several parameters. *Behavioral Sciences*, **14**, 81.

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

CORSO, J. F. (1963). A theoretico-historical review of the threshold concept. *Psychological Bulletin*, **60**, 356-370.

COX, D. R., & SNELL, E. J. (1989). *Analysis of binary data*. London: Chapman & Hall.

EMERSON, P. L. (1986). A quadrature method for Bayesian sequential threshold estimation. *Perception & Psychophysics*, **39**, 381-383.

FECHNER, G. T. (1965). *Revision der Hauptpuncte der Psychophysik* [Revision of the main topics of psychophysics]. Amsterdam: Bonset. (Original work published 1882)

FECHNER, G. T. (1966). In D. H. Howes & E. C. Boring (Eds.), and H. E. Adler (Trans.), *Elements of psychophysics*. Holt, Rinehart & Winston. (Original work published 1860)

FINNEY, D. J. (1971). *Probit analysis*. Cambridge: Cambridge University Press.

FISHER, R. A. (1912). On the absolute criterium for fitting frequency curves. *Messenger of Mathematics*, **41**, 155-160.

FLETCHER, R. (1995). *Practical methods of optimization*. Chichester, U.K.: Wiley.

GARCÍA-PÉREZ, M. A. (1998). Forced-choice staircases: Some little known facts. *Vision Research*, **38**, 1861-1881.

GREEN, D. M. (1990). Stimulus selection in adaptive psychophysical procedures. *Journal of the Acoustical Society of America*, **87**, 2662-2674.

GREEN, D. M. (1993). A maximum-likelihood method for estimating thresholds in a yes–no task. *Journal of the Acoustical Society of America*, **87**, 2096-2105.

GREEN, D. M., & SWETS, J. A. (1966). *Signal detection theory and psychophysics*. Los Altos, CA: Peninsula.

GUILFORD, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.

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

HAMBLETON, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (pp. 147-200). New York: Macmillan.

HAMBLETON, R. K., & SWAMINATHAN, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer Academic Publications.

HARVEY, L. O., JR. (1986). Efficient estimation of sensory thresholds. *Behavior Research Methods, Instruments, & Computers*, **18**, 623-632.

HARVEY, L. O., JR. (1997). Efficient estimation of sensory thresholds with ML-PEST. *Spatial Vision*, **11**, 121-128.

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

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. (1994). Comparison of the QUEST and related methods for measuring thresholds: Efficiency, bias and practical considerations. *Vision Research*, **34**, 885-912.

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

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

MACMILLAN, N. A., & CREELMAN, D. C. (1991). *Detection theory: A user'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.

MARTZ, H. F., & WALLER, R. A. (1982). *Bayesian reliability analysis*. New York: Wiley.

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.

MORGAN, B. J. T. (1992). *Analysis of quantal response data*. London: Chapman & Hall.

MÜLLER, G. E. (1879). Über die Maßbestimmungen des Ortssinnes der Haut mittelst der Methode der richtigen und falschen Fälle [On mea-

suring the spatial sense of the skin with the method of right and wrong cases]. *Pflügers Archiv*, **19**, 191-235.

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

NELDER, J. A., & MEAD, R. (1965). A simplex method for function minimization. *Computer Journal*, **7**, 308-313.

OLSSON, D. M., & NELSON, L. S. (1975). The nelder and mead simplex procedure for function minimization. *Technometrics*, **17**, 45-51.

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.

PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W. T., & FLANNERY, B. P. (1992). *Numerical recipes in C: The art of scientific computing* (2nd ed.). 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.

SWANSON, W. H., & BIRCH, E. E. (1992). Extracting thresholds from noisy psychophysical data. *Perception & Psychophysics*, **51**, 409-422.

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

TREUTWEIN, B. (1997). YAAP: Yet another adaptive procedure. *Spatial Vision*, **11**, 129-134.

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

TUKEY, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

URBAN, F. M. (1908). *The application of statistical methods to problems of psychophysics*. Philadelphia: Psychological Clinic Press.

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

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

WETHERILL, G. B. (1963). Sequential estimation of quantal response curves. *Journal of the Royal Statistical Society*, **25B**, 1-48.

NOTES

1. In the bioassay literature, the high and low asymptotes are usually called *natural immunity* and *natural response rate*, instead of lapsing and guessing rate.

2. This assumption of independence has sometimes been named the *high threshold assumption* (Green & Swets, 1966).

3. The routines have been implemented in the programming languages Modula-2 and are currently maintained in Component Pascal (see http://www.oberon.ch). Readers interested in the program source code or a machine-translated version in C should contact the first author.

4. We thank an anonymous reviewer for drawing our attention to this fact. Our approach enhances such a likelihood maximization by including Bayesian constraints. All available implementations of probit analysis, to our knowledge, stay with Finney's earlier work and use the iteratively reweighted linear regression on the transformed data.

5. For the cumulative normal, the position parameter θ is the mean, and the spread parameter σ is the standard deviation of the normal probability density function.

6. The Gumbel distribution plotted over a linear abscissa is equivalent to the Weibull plotted over a logarithmic one.

7. For an overview of available optimization routines, see Fletcher (1995) or take a look at the Web page http://www.mcs.anl.gov/home/otc/Guide

8. Although actually $\mathcal{L}'$ was minimized, we speak of maximizing the likelihood $\mathcal{L}$.

9. Using a rectangular distribution as prior for a specific parameter would be equivalent to pure ML estimation and allows someone who is extremely skeptical about Bayesian priors to resort to ML.

10. A second anonymous reviewer drew our attention to the similarity between our approach of fitting a psychometric function and the Bayesian estimation methods used to fit item-characteristic curves in item response theory (IRT; see, for example, Hambleton, 1989; Hambleton & Swaminathan, 1985). In IRT, as in psychophysics, the parameters of a sigmoid function are estimated and data are also scarce. In psychophysics, the independent variable is the stimulus level (a physical quantity), whereas, in IRT, the independent variable is the (unobservable) ability of a subject that underlies her/his performance in a specific test item—that is, a variable that can be scaled arbitrarily.

11. We used a linear congruential random number generator with seed 65539, multiplier 69069, and modulus 2^{31}, which results in a period length of approximately 2^{29} (Routine RN32 from the CERN library).

12. It has been reported that normal human subjects are unable to behave randomly (Brugger, Landis, & Regard, 1990).

(Manuscript received August 27, 1996;
revision accepted for publication December 18, 1997.)