Perception & Psychophysics
2001, 63 (8), 1293-1313

The psychometric function: I. Fitting, sampling, and goodness of fit

FELIX A. WICHMANN and N. JEREMY HILL
University of Oxford, Oxford, England

The psychometric function relates an observer's performance to an independent variable, usually some physical quantity of a stimulus in a psychophysical task. This paper, together with its companion paper (Wichmann & Hill, 2001), describes an integrated approach to (1) fitting psychometric functions, (2) assessing the goodness of fit, and (3) providing confidence intervals for the function's parameters and other estimates derived from them, for the purposes of hypothesis testing. The present paper deals with the first two topics, describing a constrained maximum-likelihood method of parameter estimation and developing several goodness-of-fit tests. Using Monte Carlo simulations, we deal with two specific difficulties that arise when fitting functions to psychophysical data. First, we note that human observers are prone to stimulus-independent errors (or *lapses*). We show that failure to account for this can lead to serious biases in estimates of the psychometric function's parameters and illustrate how the problem may be overcome. Second, we note that psychophysical data sets are usually rather small by the standards required by most of the commonly applied statistical tests. We demonstrate the potential errors of applying traditional χ^2 methods to psychophysical data and advocate use of Monte Carlo resampling techniques that do not rely on asymptotic theory. We have made available the software to implement our methods.

The performance of an observer on a psychophysical task is typically summarized by reporting one or more *response thresholds*—stimulus intensities required to produce a given level of performance—and by characterization of the rate at which performance improves with increasing stimulus intensity. These measures are derived from a *psychometric function*, which describes the dependence of an observer's performance on some physical aspect of the stimulus: One example might be the relation between the contrast of a visual stimulus and the observer's ability to detect it.

Fitting psychometric functions is a variant of the more general problem of modeling data. Modeling data is a three-step process. First, a model is chosen, and the parameters are adjusted to minimize the appropriate error metric or loss function. Second, error estimates of the parameters are derived, and third, the goodness of fit between the model and the data is assessed. This paper is concerned with the first and the third of these steps, parameter estimation and goodness-of-fit assessment. Our companion paper (Wichmann & Hill, 2001) deals with the second step and illustrates how to derive reliable error estimates on the fitted parameters. Together, the two papers provide an integrated approach to fitting psychometric functions, evaluating goodness of fit, and obtaining confidence intervals for parameters, thresholds, and slopes, avoiding the known sources of potential error.

This paper is divided into two major subsections, *fitting psychometric functions* and *goodness of fit*. Each subsection itself is again subdivided into two main parts: first, an introduction to the issue, and second, a set of simulations addressing the issue raised in the respective introduction.

NOTATION

We adhere mainly to the typographic conventions frequently encountered in statistical texts (Collett, 1991; Dobson, 1990; Efron & Tibshirani, 1993). Variables are denoted by uppercase italic letters, and observed values are denoted by the corresponding lowercase letters—for example, y is a realization of the random variable Y. Greek letters are used for parameters, and a circumflex for estimates; thus, parameter β is estimated by $\hat{\beta}$. Vectors are denoted by boldface lowercase letters, and matrices by boldface italic uppercase letters. The ith element of a vector $\mathbf{x}$ is denoted by x_i. The probability density function of the random variable Y (or the probability distribution if Y is discrete) with $\boldsymbol{\theta}$ as the vector of parameters of the distribution is written as $p(y;\boldsymbol{\theta})$. Simulated data sets (replications) are indicated by an asterisk—for example, $\hat{\alpha}_i{}^*$ is

This research was supported by a Wellcome Trust Mathematical Biology Studentship, a Jubilee Scholarship from St. Hugh's College, Oxford, and a Fellowship by Examination from Magdalen College, Oxford, to F.A.W. and by a grant from the Christopher Welch Trust Fund and a Maplethorpe Scholarship from St. Hugh's College, Oxford, to N.J.H. We are indebted to Andrew Derrington, Karl Gegenfurtner, and Bruce Henning for helpful comments and suggestions. This paper benefited considerably from conscientious peer review, and we thank our reviewers David Foster, Stan Klein, Marjorie Leek, and Bernhard Treutwein for helping us improve our manuscript. Software implementing the methods described in this paper is available written in MATLAB; contact F.A.W. at the address provided or see http://users.ox.ac.uk/~sruoxfor/psychofit/. Correspondence concerning this article should be addressed to F. A. Wichmann, Max-Planck-Institut für Biologische Kybernetik, Spemannstraße 38, D-72076 Tübingen, Germany (e-mail: felix@tuebingen. mpg.de).

the value of $\hat{\alpha}$ in the ith Monte Carlo simulation. The nth quantile of a distribution $\mathbf{x}$ is denoted by $\mathbf{x}^{(n)}$.

FITTING PSYCHOMETRIC FUNCTIONS

Background

To determine a threshold, it is common practice to fit a two-parameter function to the data and to compute the inverse of that function for the desired performance level. The slope of the fitted function at a given level of performance serves as a measure of the change in performance with changing stimulus intensity. Statistical estimation of parameters is routine in data modeling (Dobson, 1990; McCullagh & Nelder, 1989): In the context of fitting psychometric functions, probit analysis (Finney, 1952 , 1971) and a maximum-likelihood search method described by Watson (1979) are most commonly employed. Recently, Treutwein and Strasburger (1999) have described a constrained generalized maximum-likelihood procedure that is similar in some respects to the method we advocate in this paper.

In the following, we review the application of maximum-likelihood estimators in fitting psychometric functions and the use of Bayesian priors in constraining the fit according to the assumptions of one's model. In particular, we illustrate how an often disregarded feature of psychophysical data—namely, the fact that observers sometimes make stimulus-independent *lapses*—can introduce significant biases into the parameter estimates. The adverse effect of nonstationary observer behavior (of which lapses are an example) on maximum-likelihood parameter estimates has been noted previously (Harvey, 1986; Swanson & Birch, 1992; Treutwein, 1995; cf. Treutwein & Strasburger, 1999). We show that the biases depend heavily on the *sampling scheme* chosen (by which we mean the pattern of stimulus values at which samples are taken) but that it can be corrected, at minimal cost in terms of parameter variability, by the introduction of an additional free but highly constrained parameter determining, in effect, the upper bound of the psychometric function.

The psychometric function. Psychophysical data are taken by sampling an observer's performance on a psychophysical task at a number of different stimulus levels. In the method of constant stimuli, each sample point is taken in the form of a block of experimental trials at the same stimulus level. In this paper, K denotes the number of such blocks or datapoints. A data set can thus be described by three vectors, each of length K: $\mathbf{x}$ will denote the stimulus levels or intensities of the blocks, $\mathbf{n}$ the number of trials or observations in each block, and $\mathbf{y}$ the observer's performance, expressed as a proportion of correct responses (in forced-choice paradigms) or positive responses (in single-interval or yes/no experiments). We will use N to refer to the total number of experimental trials, $N = \sum n_i$.

To model the process underlying experimental data, it is common to assume the number of correct responses $y_i n_i$ in a given block i to be the sum of random samples from a Bernoulli process with a probability of success p_i. A model must then provide a psychometric function $\psi(x)$, which specifies the relationship between the underlying probability of a correct (or positive) response p and the stimulus intensity x. A frequently used general form is

$$\psi(x;\alpha,\beta,\gamma,\lambda) = \gamma + (1-\gamma-\lambda)F(x;\alpha,\beta). \qquad (1)$$

The shape of the curve is determined by the parameters $\{\alpha, \beta, \gamma, \lambda\}$, to which we shall refer collectively by using the parameter vector $\boldsymbol{\theta}$, and by the choice of a two-parameter function F, which is typically a sigmoid function, such as the Weibull, logistic, cumulative Gaussian, or Gumbel distribution.[1]

The function F is usually chosen to have range [0, 1], [0, 1), or (0,1). Thus, the parameter γ gives the lower bound of $\psi(x;\boldsymbol{\theta})$, which can be interpreted as the base rate of performance in the absence of a signal. In forced-choice paradigms (n-AFC), this will usually be fixed at the reciprocal of the number of alternatives per trial. In yes/no paradigms, it is often taken as corresponding to the *guess rate*, which will depend on the observer and experimental conditions. In this paper, we will use examples from only the 2AFC paradigm, and thus assume γ to be fixed at .5. The upper bound of the curve—that is, the performance level for an arbitrarily large stimulus signal—is given by $1 - \lambda$. For yes/no experiments, λ corresponds to the *miss rate*, and in n-AFC experiments, it is, similarly, a reflection of the rate at which observers *lapse*, responding incorrectly regardless of stimulus intensity.[2] Between the two bounds, the shape of the curve is determined by α and β. The exact meaning of α and β depends on the form of the function chosen for F, but together they will determine two independent attributes of the psychometric function: its displacement along the abscissa and its slope.

We shall assume that F describes the performance of the underlying psychological mechanism of interest. Altough it is important to have correct values for γ and λ, the values themselves are of secondary scientific interest, since they arise from the stimulus-independent mechanisms of guessing and lapsing. Therefore, when we refer to the *threshold* and *slope* of a psychometric function, we mean the inverse of F at some particular performance level as a measure of displacement and the gradient of F at that point as a measure of slope. Where we do not specify a performance level, the value .5 should be assumed: Thus *threshold* refers $F^{-1}_{0.5}$ to and *slope* refers to F' evaluated at $F^{-1}_{0.5}$. In our 2AFC examples, these values will roughly correspond to the stimulus value and slope at the 75% correct point, although the exact predicted performance will be affected slightly by the (small) value of λ.

Maximum-likelihood estimation. Likelihood maximization is a frequently used technique for parameter estimation (Collett, 1991; Dobson, 1990; McCullagh & Nelder, 1989). For our problem, provided that the values of $\mathbf{y}$ are assumed to have been generated by Bernoulli processes, it is straightforward to compute a likelihood value for a particular set of parameters $\boldsymbol{\theta}$, given the observed values $\mathbf{y}$. The likelihood function $L(\boldsymbol{\theta};\mathbf{y})$ is the same as the probability function $p(\mathbf{y}|\boldsymbol{\theta})$ (i.e., the probability of having obtained data $\mathbf{y}$ given hypothetical generating parameters $\boldsymbol{\theta}$)—note, however, the reversal of order in the notation, to stress that once the data have been gathered, $\mathbf{y}$ is fixed, and $\boldsymbol{\theta}$ is the variable. The maximum-likelihood estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is simply that set of parameters for which the likelihood value is largest: $L(\hat{\boldsymbol{\theta}};\mathbf{y}) \geq L(\boldsymbol{\theta};\mathbf{y})$ for all $\boldsymbol{\theta}$. Since the logarithm is a monotonic function, the log-likelihood function $l(\boldsymbol{\theta};\mathbf{y}) = l(\boldsymbol{\theta};\mathbf{y})$ is also maximized by the same estimator $\hat{\boldsymbol{\theta}}$, and this frequently proves to be easier to maximize numerically. For our situation, it is given by

$$l(\theta; y) = \sum_{i=1}^{K} \log\begin{pmatrix} n_i \\ y_i n_i \end{pmatrix} + y_i n_i \log \psi(x_i;\theta) + \left((1-y_i\right)n_i \log\left[1-\psi\left(x_i;\theta\right)\right]. \quad (2)$$

In principle, $\hat{\boldsymbol{\theta}}$ can be found by solving for the points at which the derivative of $l(\boldsymbol{\theta};\mathbf{y})$ with respect to all the parameters is zero: This gives a set of local minima and maxima, from which the global maximum of $l(\hat{\boldsymbol{\theta}};\mathbf{y})$ is selected. For most practical applications, however, $\hat{\boldsymbol{\theta}}$ is determined by iterative methods, maximizing those terms of Equation 2 that depend on $\boldsymbol{\theta}$. Our implementation of log-likelihood maximization uses the multidimensional Nelder–Mead simplex search algorithm,[3] a description of which can be found in chapter 10 of Press, Teukolsky, Vetterling, and Flannery (1992).

Bayesian priors. It is sometimes possible that the maximum-likelihood estimate $\hat{\boldsymbol{\theta}}$ contains parameter values that are either nonsensical or inappropriate. For example, it can happen that the best fit to a particular data set has a negative value for λ, which is uninterpretable as a lapse rate and implies that an observer's performance can exceed 100% correct—clearly nonsensical psychologically, even though $l(\boldsymbol{\theta}; \mathbf{y})$ may be a real value for the particular stimulus values in the data set.

It may also happen that the data are best fit by a parameter set containing a large λ (greater than .06, for example). A large λ is interpreted to mean that the observer makes a large proportion of incorrect responses no matter how great the stimulus intensity—in most normal psychophysical situations, this means that the experiment was not performed properly and that the data are invalid. If the observer genuinely has a lapse rate greater than .06, he or she requires extra encouragement or, possibly, replacement. However, misleadingly large λ values may also be fitted when the observer performs well, but there are no samples at high performance values.

In both cases, it would be better for the fitting algorithm to return parameter vectors that may have a lower log-likelihood than the global maximum but that contain more realistic values. Bayesian priors provide a mechanism for constraining parameters within realistic ranges, based on the experimenter's prior beliefs about the likelihood of particular values. A prior is simply a relative probability distribution $W(\boldsymbol{\theta})$, specified in advance, which weights the likelihood calculation during fitting: The fitting process therefore maximizes $W(\boldsymbol{\theta})L(\boldsymbol{\theta};\mathbf{y})$ or $\log W(\boldsymbol{\theta}) + l(\boldsymbol{\theta};\mathbf{y})$, instead of the unweighted metrics.

The exact form of $W(\boldsymbol{\theta})$ is to be chosen by the experimenter, given the experimental context. The ideal choice for $W(\lambda)$ would be the distribution of rates of stimulus-independent error for the current observer on the current task. Generally, however, one has not enough data to estimate this distribution. For the simulations reported in this paper, we chose $W(\lambda) = 1$ for $0 \leq \lambda \leq .06$; otherwise, $W(\lambda) = 0$[4]—that is, we set a limit of .06 on λ, and weight smaller values equally with a flat prior.[5,6] For data analysis, we generally do not constrain the other parameters, except to limit them to values for which $\psi(x;\boldsymbol{\theta})$ is real.

Avoiding bias caused by observers' lapses. In studies in which sigmoid functions are fitted to psychophysical data, particularly where the data come from forced-choice paradigms, it is common for experimenters to fix $\lambda = 0$, so that the upper bound of $\psi(x;\boldsymbol{\theta})$ is always 1.0. Thus, it is assumed that observers make no stimulus-independent errors. Unfortunately, maximum-likelihood parameter estimation as described above is extremely sensitive to such stimulus-independent errors, with a consequent bias in threshold and slope estimates (Harvey, 1986; Swanson & Birch, 1992).

Figure 1 illustrates the problem. The dark circles indicate the proportion of correct responses made by an observer in six blocks of trials in a 2AFC visual detection task. Each datapoint represents 50 trials, except for the last one, at stimulus value 3.5, which represents 49 trials: The observer still has one trial to perform to complete the block. If we were to stop here and fit a Weibull function to the data, we would obtain the curve plotted as a dark solid line. Whether or not λ is fixed at 0 during the fit, the maximum-likelihood parameter estimates are the same: $\{\hat{\alpha} = 1.573, \hat{\beta} = 4.360, \hat{\lambda} = 0\}$. Now suppose that, on the 50th trial of the last block, the observer blinks and misses the stimulus, is consequently forced to guess, and happens to guess wrongly. The new position of the datapoint at stimulus value 3.5 is shown by the light triangle: It has dropped from 1.00 to .98 proportion correct.

The solid light curve shows the results of fitting a two-parameter psychometric function (i.e., allowing α and β to vary, but keeping λ fixed at 0). The new fitted parameters are $\{\hat{\alpha} = 2.604, \hat{\beta} = 2.191\}$. Note that the slope of the fitted function has dropped dramatically in the space of one trial—in fact, from a value of 1.045 to 0.560. If we allow λ to vary in our new fit, however, the effect on parameters is slight—$\{\hat{\alpha} = 1.543, \hat{\beta} = 4.347, \hat{\lambda} = .014\}$—and thus, there is little change in slope: $\mathrm{d}\hat{F}/\mathrm{d}x$ evaluated at $x = F_{0.5}^{-1}$ is 1.062.

The misestimation of parameters shown in Figure 1 is a direct consequence of the binomial log-likelihood error metric because of its sensitivity to errors at high levels of predicted performance:[7] since $\psi(x;\boldsymbol{\theta}) \rightarrow 1$, so, in the third term of Equation 2, $(1 - y_i)n_i \log[1 - \psi(x_i;\boldsymbol{\theta})] \rightarrow -\infty$ un-

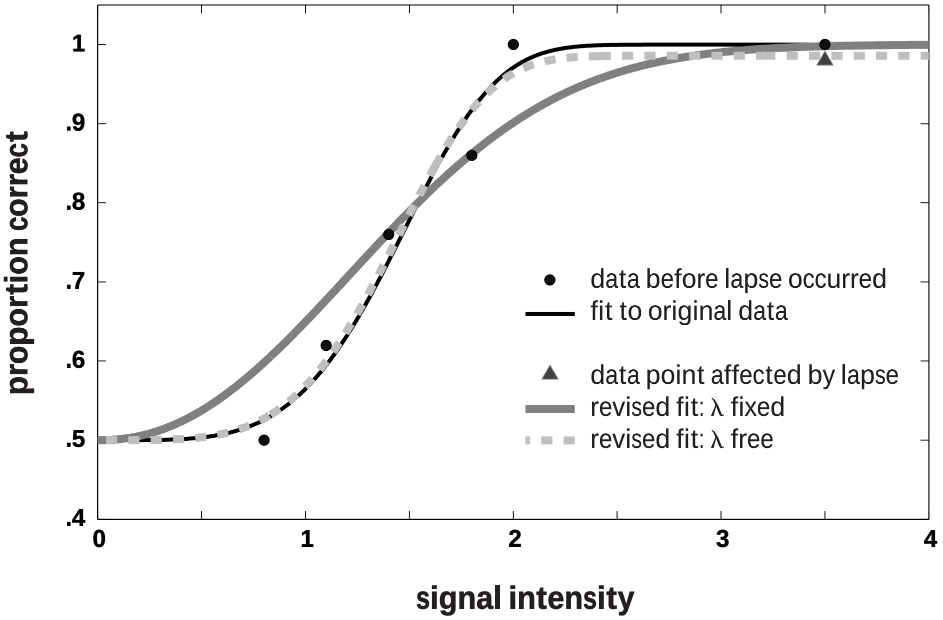

Figure 1. Dark circles show data from a hypothetical observer prior to lapsing. The solid dark line is a maximum-likelihood Weibull fit to the data set. The triangle shows a datapoint after the observer lapsed once during a 50-trial block. The solid light line shows the (poor) traditional two-parameter Weibull fit with λ fixed; the broken light line shows the suggested three-parameter Weibull fit with λ free to vary.

less the coefficient $(1 - y_i)n_i$ is 0. Since y_i represents observed proportion correct, the coefficient is 0 as long as performance is perfect. However, as soon as the observer lapses, the coefficient becomes nonzero and allows the large negative log term to influence the log-likelihood sum, reflecting the fact that observed proportions less than 1 are extremely unlikely to have been generated from an expected value that is very close to 1. Log-likelihood can be raised by lowering the predicted value at the last stimulus value, $\psi(3.5,\boldsymbol{\theta})$. Given that λ is fixed at 0, the upper asymptote is fixed at 1.0; hence, the best the fitting algorithm can do in our example to lower $\psi(3.5,\boldsymbol{\theta})$ is to make the psychometric function shallower.

Judging the fit by eye, it does not appear to capture accurately the rate at which performance improves with stimulus intensity. (Proper Monte Carlo assessments of goodness of fit are described later in this paper.)

The problem can be cured by allowing λ to take a nonzero value, which can be interpreted to reflect our belief as experimenters that observers can lapse and that, therefore, in some cases, their performance might fall below 100% despite arbitrarily large stimulus values. To obtain the optimum value of λ and, hence, the most accurate estimates for the other parameters, we allow λ to vary in the maximum-likelihood search. However, it is constrained within the narrow range [0,.06], reflecting our beliefs concerning its likely values[8] (see the previous section, on Bayesian priors).

The example of Figure 1 might appear exaggerated; the distortion in slope was obtained by placing the last sample point (at which the lapse occurred) at a comparatively high stimulus value relative to the rest of the data set. The question remains: How serious are the consequences of assuming a fixed λ for sampling schemes one might readily employ in psychophysical research?

Simulations

To test this, we conducted Monte Carlo simulations; six-point data sets were generated binomially assuming a 2AFC design and using a standard underlying performance function $F(x;\{\alpha_{\text{gen}}, \beta_{\text{gen}}\})$, which was a Weibull function with parameters $\alpha_{\text{gen}} = 10$ and $\beta_{\text{gen}} = 3$. Seven different sampling schemes were used, each dictating a different distribution of datapoints along the stimulus axis. They are shown in Figure 2: Each horizontal chain of symbols represents one of the schemes, marking the stimulus values at which the six sample points are placed. The different symbol shapes will be used to identify the sampling schemes in our results plots. To provide a frame of reference, the solid curve shows $0.5 + 0.5(F(x;\{\alpha_{\text{gen}}, \beta_{\text{gen}}\})$, with the 55%, 75%, and 95% performance levels marked by dotted lines.

Our seven schemes were designed to represent a range of different sampling distributions that could arise during "everyday" psychophysical laboratory practice, including those skewed toward low performance values (s4) or high performance values (s3 and s7), those that are clustered around threshold (s1), those that are spread out away from the threshold (s5), and those that span the range from 55% to 95% correct (s2). As we shall see, even for a fixed num-

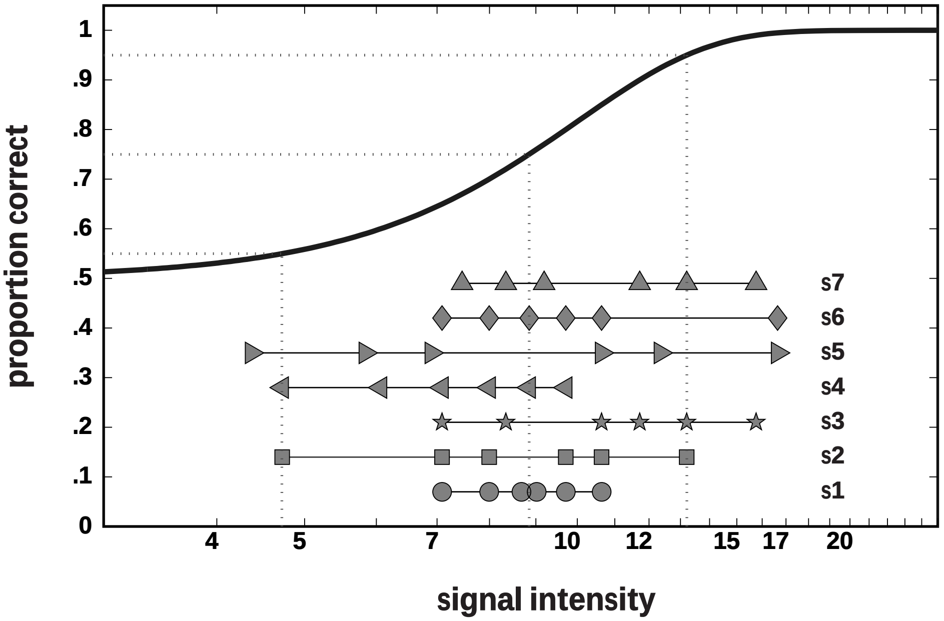

Figure 2. A two-alternative forced-choice Weibull psychometric function with parameter vector θ = {10, 3, .5, 0} on semilogarithmic coordinates. The rows of symbols below the curve mark the *x* values of the seven different sampling schemes used throughout the remainder of the paper.

ber of sample points and a fixed number of trials per point, biases in parameter estimation and goodness-of-fit assessment (this paper) as well as the width of confidence intervals (companion paper, Wichmann & Hill, 2001), all depend heavily on the distribution of stimulus values **x**.

The number of datapoints was always 6, but the number of observations per point was 20, 40, 80, or 160. This meant that the total number of observations N could be 120, 240, 480, or 960.

We also varied the rate at which our simulated observer lapsed. Our model for the processes involved in a single trial was as follows: For every trial at stimulus value x_i, the observer's probability of correct response $\psi_{\text{gen}}(x_i)$ is given by $0.5 + 0.5F(x_i;\{\alpha_{\text{gen}}, \beta_{\text{gen}}\})$, except that there is a certain small probability that something goes wrong, in which case $\psi_{\text{gen}}(x_i)$ is instead set at a constant value k. The value of k would depend on exactly *what* goes wrong. Perhaps the observer suffers a loss of attention, misses the stimulus, and is forced to guess; then, k would be .5, reflecting the probability of the guess' being correct. Alternatively, lapses might occur because the observer fails to respond within a specified response interval, which the experimenter interprets as an incorrect response, in which case $k = 0$. Or perhaps k has an intermediate value that reflects a probabilistic combination of these two events and/or other potential mishaps. In any case, provided we assume that such events are independent, that their probability of occurrence is constant throughout a single block of trials, and that k is constant, the simulated observer's overall performance on a block is binomially distributed, with an underlying probability that can be expressed with Equation 1—it is easy to show that variations in k or in the probability of a mishap are described by a change in λ. We shall use λ_{gen} to denote the generating function's λ parameter, possible values for which were 0, .01, .02, .03, .04, and .05.

To generate each data set, then, we chose (1) a sampling scheme, which gave us the vector of stimulus values **x**, (2) a value for N, which was divided equally among the elements of the vector **n** denoting the number of observations in each block, and (3) a value for λ_{gen}, which was assumed to have the same value for all blocks of the data set.[9] We then obtained the simulated performance vector **y**: For each block i, the proportion of correct responses y_i was obtained by finding the proportion of a set of n_i random numbers[10] that were less than or equal to $\psi_{\text{gen}}(x_i)$. A maximum-likelihood fit was performed on the data set described by **x**, **y**, and **n**, to obtain estimated parameters $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\lambda}$. There were two fitting regimes: one in which $\hat{\lambda}$ was fixed at 0, and one in which it was allowed to vary but was constrained within the range [0, .06]. Thus, there were 336 conditions: 7 sampling schemes × 6 values for λ_{gen} × 4 values for N × 2 fitting regimes. Each condition was replicated 2,000 times, using a new randomly generated data set, for a total of 672,000 simulations overall, requiring 3.024×10^8 simulated 2AFC trials.

Simulation results: 1. Accuracy. We are interested in measuring the accuracy of the two fitting regimes in estimating the threshold and slope, whose true values are given by the stimulus value and gradient of $F(x;\{\alpha_{\text{gen}}, \beta_{\text{gen}}\})$ at the point at which $F(x;\{\alpha_{\text{gen}}, \beta_{\text{gen}}\}) = 0.5$. The true values are 8.85 and 0.118, respectively—that is, it is

$F(8.85;\{10,3\}) = 0.5$ and $F'(8.85;\{10,3\}) = 0.118$. From each simulation, we use the estimated parameters to obtain the threshold and slope of $F(x;\{\hat{\alpha}, \hat{\beta}\})$. The medians of the distributions of 2,000 thresholds and slopes from each condition are plotted in Figure 3. The left-hand column plots median estimated threshold, and the right-hand column plots median estimated slope, both as a function of λ_{gen}. The four rows correspond to the four values of *N*: The total number of observations increases down the page. Symbol shape denotes sampling scheme as per Figure 2. Light symbols show the results of fixing λ at 0, and dark symbols show the results of allowing λ to vary during the fitting process. The true threshold and slope values (i.e., those obtained from the generating function) are shown by the solid horizontal lines.

The first thing to notice about Figure 3 is that, in all the plots, there is an increasing bias in some of the sampling schemes' median estimates as λ_{gen} increases. Some sampling schemes are relatively unaffected by the bias. By using the shape of the symbols to refer back to Figure 2, it can be seen that the schemes that are affected to the greatest extent (s3, s5, s6, and s7) are those containing sample points at which $F(x,\{\alpha_{\text{gen}}, \beta_{\text{gen}}\}) > 0.9$, whereas the others (s1, s2, and s4) contain no such points and are affected to a lesser degree. This is not surprising, bearing in mind the foregoing discussion of Equation 1: Bias is most likely where high performance is expected.

Generally, then, the variable-λ regime performs better than the fixed-λ regime in terms of bias. The one exception to this can be seen in the plot of median slope estimates for $N = 120$ (20 observations per point): Here, there is a slight *upward* bias in the variable-λ estimates, an effect that varies according to sampling scheme but that is relatively unaffected by the value of λ_{gen}. In fact, for $\lambda_{\text{gen}} \leq .02$, the downward bias from the fixed-λ fits is smaller, or at least no larger, than the upward bias from fits with variable λ. Note, however, that an increase in *N* to 240 or more improves the variable-λ estimates, reducing the bias and bringing the medians from the different sampling schemes together. The variable-λ fitting scheme is essentially unbiased for $N \geq 480$, independent of the sampling scheme chosen. By contrast, the value of *N* appears to have little or no effect on the absolute size of the bias inherent in the susceptible fixed-λ schemes.

Simulation results: 2. Precision. In Figure 3, the bias seems fairly small for threshold measurements (maximally, about 8% of our stimulus range when the fixed-λ fitting regime is used or about 4% when λ is allowed to vary). For slopes, only the fixed-λ regime is affected, but the effect, expressed as a percentage, is more pronounced (up to 30% underestimation of gradient).

However, note that however large or small the bias appears when expressed in terms of stimulus units, knowledge about an estimator's *precision* is required in order to assess the severity of the bias. Severity in this case means the extent to which our estimation procedure leads us to make errors in hypothesis testing: finding differences between experimental conditions where none exist (Type I errors) or failing to find them when they do exist (Type II). The bias of an estimator must thus be evaluated relative to its variability. A frequently applied rule of thumb is that a good estimator should be biased by less than 25% of its standard deviation (Efron & Tibshirani, 1993).

The variability of estimates in the context of fitting psychometric functions is the topic of our companion paper (Wichmann & Hill, 2001), in which we shall see that one's chosen sampling scheme and the value of *N* both have a profound effect on confidence interval width. For now, and without going into too much detail, we are merely interested in knowing how our decision to employ a fixed-λ or a variable-λ regime affects variability (precision) for the various sampling schemes and to use this information to assess the severity of bias.

Figure 4 shows two illustrative cases, plotting results for two contrasting schemes at $N = 480$. The upper pair of plots shows the s7 sampling scheme, which, as we have seen, is highly susceptible to bias when λ is fixed at 0 and observers lapse. The lower pair shows s1, which we have already found to be comparatively resistant to bias. As before, thresholds are plotted on the left and slopes on the right, light symbols represent the fixed-λ fitting regime, dark symbols represent the variable-λ fitting regime, and the true values are again shown as solid horizontal lines. Each symbol's position represents the median estimate from the 2,000 fits at that point, so they are exactly the same as the upward triangles and circles in the $N = 480$ plots of Figure 3. The vertical bars show the interval between the 16th and the 84th percentiles of each distribution (these limits were chosen because they give an interval with coverage of .68, which would be approximately the same as the mean plus or minus one standard deviation (*SD*) if the distributions were Gaussian). We shall use WCI_{68} as shorthand for *width of the 68% confidence interval* in the following.

Applying the rule of thumb mentioned above (bias ≤ 0.25 *SD*), bias in the fixed-λ condition in the threshold estimate is significant for $\lambda_{\text{gen}} > .02$, for both sampling schemes. The slope estimate of the fixed-λ fitting regime and the sampling scheme s7 is significantly biased once observers lapse—that is, for $\lambda_{\text{gen}} \geq .01$. It is interesting to see that even the threshold estimates of s1, with all sample points at $p < .8$, are significantly biased by lapses. The slight bias found with the variable-λ fitting regime, however, is not significant in any of the cases studied.

We can expect the variance of the distribution of estimates to be a function of *N*—the WCI_{68} gets smaller with increasing *N*. Given that the absolute magnitude of bias stays the same for the fixed-λ fitting regime, however, bias will become more problematic with increasing *N*: For $N = 960$, the bias in threshold and slope estimates is significant for all sampling schemes and virtually all nonzero lapse rates. Frequently, the true (generating) value is not even within the 95% confidence interval. Increasing the number of observations by using the fixed-λ fitting regime, contrary to what one might expect, *increases* the likelihood

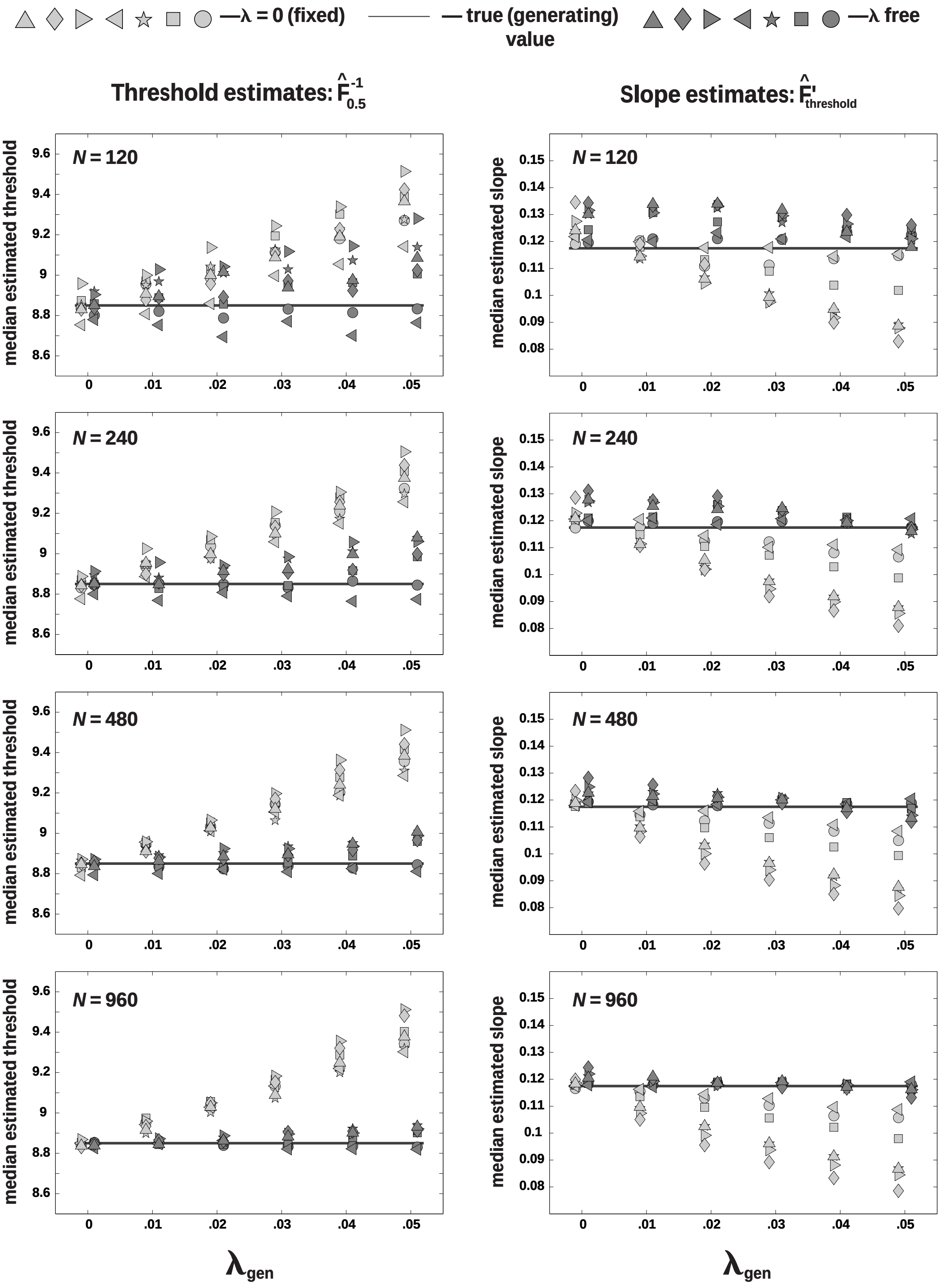

Figure 3. Median estimated thresholds and median estimated slopes are plotted in the left-hand and right-hand columns, respectively; both are shown as a function of λ_{gen}. The four rows correspond to four values of N (120, 240, 480, and 960). Symbol shapes denote the different sampling schemes (see Figure 2). Light symbols show the results of fixing λ at 0; dark symbols show the results of allowing λ to vary during the fitting. The true threshold and slope values are shown by the solid horizontal lines.

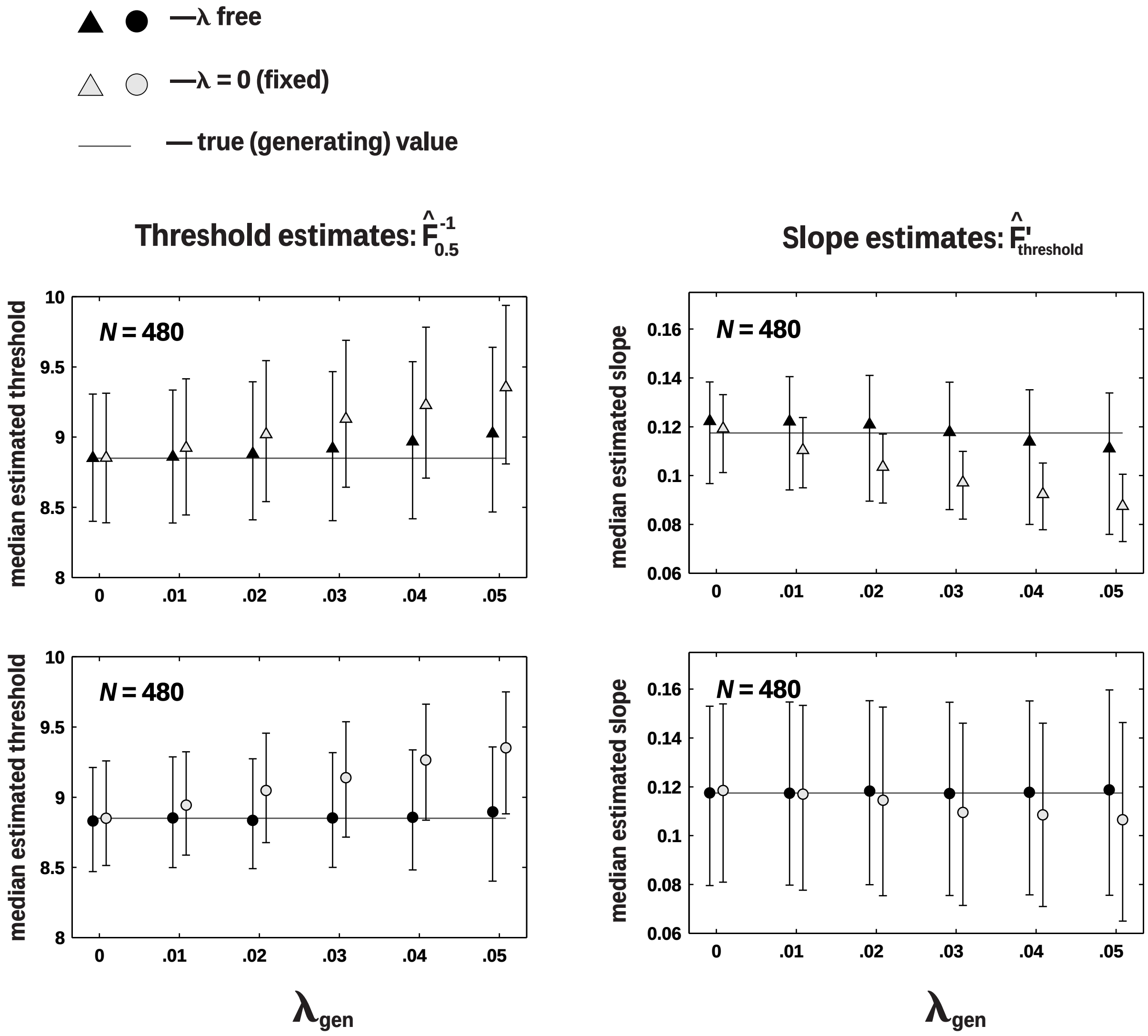

Figure 4. Median estimated thresholds and median estimated slopes are plotted as a function of λ_{gen} on the left and right, respectively. The vertical bars show WCI_{68} (see the text for details). Data for two sampling schemes, s1 and s7, are shown for N = 480.

of Type I or II errors. Again, the variable-λ fitting regime performs well: The magnitude of median bias decreases approximately in proportion to the decrease in WCI_{68}. Estimates are essentially unbiased.

For small N (i.e., $N = 120$), WCI_{68}s are larger than those shown in Figure 4. Very approximately,[11] they increase with $1/\sqrt{N}$; even for $N = 120$, bias in slope and threshold estimates is significant for most of the sampling schemes in the fixed-λ fitting regime once $\lambda_{gen} \geq .02$ or .03. The variable-λ fitting regime performs better again, although for some sampling schemes (s3, s5, and s7) bias is significant at $\lambda_{gen} \geq .04$, because bias increases disproportionally relative to the increase in WCI_{68}.

A second observation is that, in the case of s7, correcting for bias by allowing λ to vary carries with it the cost of reduced precision. However, as was discussed above, the alternative is uninviting: With λ fixed at 0, the true slope does not even lie within the 68% confidence interval of the distribution of estimates for $\lambda_{gen} \geq .02$. For s1, however, allowing λ to vary brings neither a significant benefit in terms of accuracy nor a significant penalty in terms of precision. We have found that these two contrasting cases are representative: Generally, there is nothing to lose by allowing λ to vary in those cases where it is not required in order to provide unbiased estimates.

Fixing lambda at nonzero values. In the previous analysis, we contrasted a variable-λ fitting regime with one having λ fixed at 0. Another possibility might be to fix λ at a small but nonzero value, such as .02 or .04. Here, we report Monte Carlo simulations exploring whether a fixed small value of λ overcomes the problems of bias while retaining the desirable property of (slightly) increased precision relative to the variable-λ fitting regime.

Simulations were repeated as before, except that λ was either free to vary or fixed at .01, .02, .03, .04, and .05, covering the whole range of λ_{gen}. (A total of 2,016,000 simulations, requiring 9.072×10^9 simulated 2AFC trials.)

Figure 5 shows the results of the simulations in the same format as that of Figure 3. For clarity, only the variable-λ fitting regime and λ fixed at .02 and .04 are plotted, using dark, intermediate, and light symbols, respectively. Since bias in the fixed-λ regimes is again largely independent of the number of trials N, only data corresponding to the intermediate number of trials is shown ($N = 240$ and 480). The data for the fixed-λ regimes indicate that both are simply shifted copies of each other—in fact, they are more or less merely shifted copies of the λ fixed at zero data presented in Figure 3. Not surprisingly, minimal bias is obtained for λ_{gen} corresponding to the fixed-λ value. The *zone* of insignificant bias around the fixed-λ value is small, however, only extending to, at most, $\lambda \pm .01$. Thus, fixing λ at, say, .01 provides unbiased and precise estimates of threshold and slope, provided the observer's lapse rate is within $0 \leq \lambda_{\text{gen}} \leq .02$. In our experience, this zone or range of good estimation is too narrow: One of us (F.A.W.) regularly fits psychometric functions to data from discrimination and detection experiments, and even for a single observer, λ in the variable-λ fitting regime takes on values from 0 to .05—no single fixed λ is able to provide unbiased estimates under these conditions.

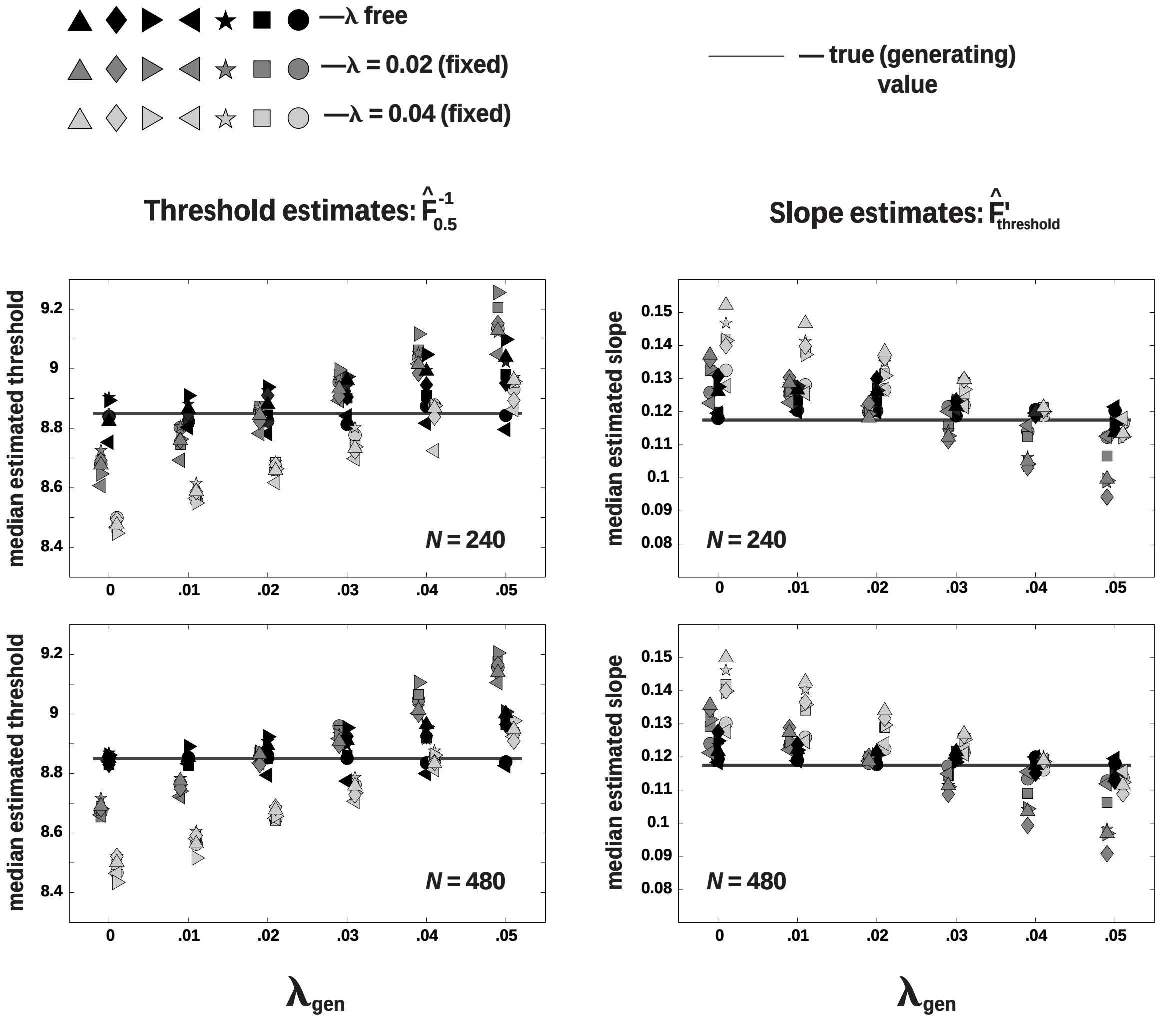

Figure 5. Data shown in the format of Figure 3; median estimated thresholds and median estimated slopes are plotted as a function of λ_{gen} in the left-hand and right-hand columns, respectively. The two rows correspond to $N = 240$ and 480. Symbol shapes denote the different sampling schemes (see Figure 2). Light symbols show the results of fixing λ at .04; medium gray symbols those for λ fixed at .02; dark symbols show the results of allowing λ to vary during the fitting. True threshold and slope values are shown by the solid horizontal lines.

Discussion and Summary

Could bias be avoided simply by choosing a sampling scheme in which performance close to 100% is not expected? Since one never knows the psychometric function exactly in advance before choosing where to sample performance, it would be difficult to avoid high performance, even if one were to want to do so. Also, there is good reason to *choose* to sample at high performance values: Precisely because data at these levels have a greater influence on the maximum-likelihood fit, they carry more information about the underlying function and thus allow more efficient estimation. Accordingly, Figure 4 shows that the precision of slope estimates is better for s7 than for s1 (cf. Lam, Mills, & Dubno, 1996). This issue is explored more fully in our companion paper (Wichmann & Hill, 2001). Finally, even for those sampling schemes that contain no sample points at performance levels above 80%, bias in threshold estimates was significant, particularly for large *N*.

Whether sampling deliberately or accidentally at high performance levels, one must allow for the possibility that observers will perform at high rates and yet occasionally lapse: Otherwise, parameter estimates may become biased when lapses occur. Thus, we recommend that varying λ as a third parameter be the method of choice for fitting psychometric functions.

Fitting a tightly constrained λ is intended as a heuristic to avoid bias in cases of nonstationary observer behavior. It is, as well, to note that the estimated parameter $\hat{\lambda}$ is, in general, *not* a very good estimator of a subject's *true* lapse rate (this was also found by Treutwein & Strasburger, 1999, and can be seen clearly in their Figures 7 and 10). Lapses are rare events, so there will only be a very small number of lapsed trials per data set. Furthermore, their directly measurable effect is small, so that only a small subset of the lapses that occur (those at high *x* values where performance is close to 100%) will affect the maximum-likelihood estimation procedure; the rest will be lost in binomial noise. With such minute effective sample sizes, it is hardly surprising that our estimates of λ per se are poor. However, we do not need to worry, because as psychophysicists we are not interested in lapses: We are interested in thresholds and slopes, which are determined by the function *F* that reflects the underlying mechanism. Therefore, we vary λ not for its own sake, but purely in order to free our threshold and slope estimates from bias. This it accomplishes well, *despite* numerically inaccurate λ estimates. In our simulations, it works well both for sampling schemes with a fixed nonzero λ_{gen} and for those with more random lapsing schemes (see note 9 or our example shown in Figure 1).

In addition, our simulations have shown that $N = 120$ appears too small a number of trials to be able to obtain reliable estimates of thresholds and slopes for some sampling schemes, even if the variable-λ fitting regime is employed. Similar conclusions were reached by O'Regan and Humbert (1989) for $N = 100$ ($K = 10$; cf. Leek, Hanna, & Marshall, 1992; McKee, Klein, & Teller, 1985). This is further supported by the analysis of bootstrap sensitivity in our companion paper (Wichmann & Hill, 2001).

GOODNESS OF FIT

Background

Assessing goodness of fit is a necessary component of any sound procedure for modeling data, and the importance of such tests cannot be stressed enough, given that fitted thresholds and slopes, as well as estimates of variability (Wichmann & Hill, 2001), are usually of very limited use if the data do not appear to have come from the hypothesized model. A common method of goodness-of-fit assessment is to calculate an error term or summary statistic, which can be shown to be asymptotically distributed according to χ^2—for example, Pearson X^2—and to compare the error term against the appropriate χ^2 distribution. A problem arises, however, since psychophysical data tend to consist of small numbers of points and it is, hence, by no means certain that such tests are accurate. A promising technique that offers a possible solution is Monte Carlo simulation, which being computationally intensive, has become practicable only in recent years with the dramatic increase in desktop computing speeds. It is potentially well suited to the analysis of psychophysical data, because its accuracy does not rely on large numbers of trials, as do methods derived from asymptotic theory (Hinkley, 1988). We show that for the typically small *K* and *N* used in psychophysical experiments, assessing goodness of fit by comparing an empirically obtained statistic against its asymptotic distribution is not always reliable: The true small-sample distribution of the statistic is often insufficiently well approximated by its asymptotic distribution. Thus, we advocate generation of the necessary distributions by Monte Carlo simulation.

Lack of fit—that is, the failure of goodness of fit—may result from failure of one or more of the assumptions of one's model. First and foremost, lack of fit between the model and the data could result from an inappropriate functional form for the model—in our case of fitting a psychometric function to a single data set, the chosen underlying function *F* is significantly different from the true one. Second, our assumption that observer responses are binomial may be false: For example, there might be serial dependencies between trials within a single block. Third, the observer's psychometric function may be nonstationary during the course of the experiment, be it due to learning or fatigue.

Usually, inappropriate models and violations of independence result in *overdispersion* or *extra-binomial variation*: "bad fits" in which datapoints are significantly further from the fitted curve than was expected. Experimenter bias in data selection (e.g., informal removal of *outliers*), on the other hand, could result in *underdispersion*: fits that are "too good to be true," in which datapoints are significantly *closer* to the fitted curve than one might expect (such data sets are reported more frequently in the psychophysical literature than one would hope[12]).

Typically, if goodness of fit of fitted psychometric functions is assessed at all, however, only overdispersion is considered. Of course, this method does not allow us to distinguish between different sources of overdispersion (wrong underlying function or violation of independence) and/or effects of learning. Furthermore, as we will show, models can be shown to be in error even if the summary statistic indicates an acceptable fit.

In the following, we describe a set of goodness-of-fit tests for psychometric functions (and parametric models in general). Most of them rely on different analyses of the residual differences between data and fit (the sum of squares of which constitutes the popular summary statistics) and on Monte Carlo generation of the statistic's distribution, against which to assess lack of fit. Finally, we show how a simple application of the *jackknife* resampling technique can be used to identify so-called *influential observations*—that is, individual points in a data set that exert undue influence on the final parameter set. Jackknife techniques can also provide an objective means of identifying outliers.

Assessing overdispersion. Summary statistics measure closeness of the data set as a whole to the fitted function. Assessing *closeness* is intimately linked to the fitting procedure itself: Selecting the appropriate error metric for fitting implies that the relevant *currency* within which to measure closeness has been identified. How good or bad the fit is should thus be assessed in the same currency.

In maximum-likelihood parameter estimation, the parameter vector $\hat{\boldsymbol{\theta}}$ returned by the fitting routine is such that $L(\hat{\boldsymbol{\theta}};\mathbf{y}) \geq L(\boldsymbol{\theta};\mathbf{y})$ for all $\boldsymbol{\theta}$. Thus, whatever error metric, Z, is used to assess goodness of fit,

$$Z(\hat{\boldsymbol{\theta}};\mathbf{y}) \geq Z(\boldsymbol{\theta};\mathbf{y}) \tag{3}$$

should hold for all $\boldsymbol{\theta}$.

Deviance. The log-likelihood ratio, or *deviance,* is a monotonic transformation of likelihood and therefore fulfills the criterion set out in Equation 3. Hence, it is commonly used in the context of generalized linear models (Collett, 1991; Dobson, 1990; McCullagh & Nelder, 1989).

Deviance, D, is defined as

$$D = 2\log\left[\frac{L(\boldsymbol{\theta}_{\max};\mathbf{y})}{L(\hat{\boldsymbol{\theta}};\mathbf{y})}\right] = 2\left[l(\boldsymbol{\theta}_{\max};\mathbf{y}) - l(\hat{\boldsymbol{\theta}};\mathbf{y})\right], \tag{4}$$

where $L(\boldsymbol{\theta}_{\max};\mathbf{y})$ denotes the likelihood of the *saturated* model—that is, a model with no residual error between empirical data and model predictions. ($\boldsymbol{\theta}_{\max}$ denotes the parameter vector such that this holds and the number of free parameters in the saturated model is equal to the total number of blocks of observations, K.) $L(\hat{\boldsymbol{\theta}};\mathbf{y})$ is the likelihood of the best-fitting model; $l(\boldsymbol{\theta}_{\max};\mathbf{y})$ and $l(\hat{\boldsymbol{\theta}};\mathbf{y})$ denote the logarithms of these quantities, respectively. Because, by definition, $l(\boldsymbol{\theta}_{\max};\mathbf{y}) \geq l(\hat{\boldsymbol{\theta}};\mathbf{y})$ for all $\boldsymbol{\theta}$, and $l(\boldsymbol{\theta}_{\max};\mathbf{y})$ is independent of $\hat{\boldsymbol{\theta}}$ (being purely a function of the data, $\mathbf{y}$), deviance fulfills the criterion set out in Equation 3. From Equation 4 we see that deviance takes values from 0 (no residual error) to infinity (observed data are impossible given model predictions).

For goodness-of-fit assessment of psychometric functions (binomial data), Equation 4 reduces to

$$D = 2\sum_{i=1}^{K}\left\{n_i y_i \log\left(\frac{y_i}{p_i}\right) + n_i\left(1-y_i\right)\log\left(\frac{1-y_i}{1-p_i}\right)\right\} \tag{5}$$

(p_i refers to the proportion correct predicted by the fitted model).

Deviance is used to assess goodness of fit, rather than likelihood or log-likelihood directly, because, for correct models, deviance for binomial data is asymptotically distributed as χ^2_K, where K denotes the number of datapoints (blocks of trials).[13] For derivation, see, for example, Dobson (1990, p. 57), McCullagh and Nelder (1989), and in particular, Collett (1991, sects. 3.8.1 and 3.8.2). Calculating D from one's fit and comparing it with the appropriate χ^2 distribution, hence, allows simple goodness-of-fit assessment, provided that the asymptotic approximation to the (unknown) distribution of deviance is accurate for one's data set. The specific values of $L(\hat{\boldsymbol{\theta}};\mathbf{y})$ or $l(\hat{\boldsymbol{\theta}};\mathbf{y})$ or by themselves, on the other hand, are less generally interpretable.

Pearson X^2. The Pearson X^2 test is widely used in goodness-of-fit assessment of multinomial data; applied to K blocks of binomial data, the statistic has the form

$$X^2 = \sum_{i=1}^{K}\frac{n_i\left(y_i - p_i\right)^2}{p_i\left(1-p_i\right)}, \tag{6}$$

with n_i, y_i, and p_i as in Equation 5. Equation 6 can be interpreted as the sum of squared residuals (each residual being given by $y_i - p_i$), standardized by their variance, $p_i(1-p_i)n_i^{-1}$. Pearson X^2 is asymptotically distributed according to χ^2 with K degrees of freedom, because the binomial distribution is asymptotically normal and χ^2_K is defined as the distribution of the sum of K squared unit-variance normal deviates. Indeed, deviance D and Pearson X^2 have the same asymptotic χ^2 distribution (but see note 13).

There are two reasons why deviance is preferable to Pearson X^2 for assessing goodness of fit after maximum-likelihood parameter estimation. First and foremost, for Pearson X^2, Equation 3 does not hold—that is, the maximum-likelihood parameter estimate $\hat{\boldsymbol{\theta}}$ will not generally correspond to the set of parameters with the smallest error in the Pearson X^2 sense—that is, Pearson X^2 errors are the wrong *currency* (see the previous section, Assessing Overdispersion). Second, differences in deviance between two models of the same family—that is, between models where one model includes terms in addition to those in the other—can be used to assess the significance of the additional free parameters. Pearson X^2, on the other hand, cannot be used for such model comparisons (Collett, 1991). This important issue will be expanded on when we introduce an objective test of outlier identification.

Simulations

As we have mentioned, both deviance for binomial data and Pearson X^2 are only asymptotically distributed according to χ^2. In the case of Pearson X^2, the approximation to the χ^2 will be reasonably good once the K individual binomial contributions to Pearson X^2 are well approximated by a normal—that is, as long as both $n_i p_i \geq 5$ and $n_i(1 - p_i) \geq 5$ (Hoel, 1984). Even for only moderately high p values like .9, this already requires n_i values of 50 or more, and $p = .98$ requires an n_i of 250.

No such simple criterion exists for deviance, however. The approximation depends not only on K and N, but importantly, on the size of the individual n_i and p_i—it is difficult to predict whether or not the approximation is *sufficiently close* for a particular data set. (see Collett, 1991, sect. 3.8.2). For binary data (i.e. , $n_i = 1$), deviance is not even asymptotically distributed according to χ^2, and for small n_i, the approximation can thus be very poor even if K is large.

Monte-Carlo-based techniques are well suited to answering any question of the kind "what distribution of values would we *expect* if . . .?" and, hence, offers a potential alternative to relying on the large-sample χ^2 approximation for assessing goodness of fit. The distribution of deviances is obtained in the following way. First, we generate B data sets $\mathbf{y}_i$*, using the best-fitting psychometric function, $\psi(x,\hat{\boldsymbol{\theta}})$, as the generating function. Then, for each of the $i = \{1, \ldots, B\}$ generated data sets $\mathbf{y}_i$*, we calculate deviance D_i*, using Equation 5, yielding the deviance distribution **D***. The distribution **D*** reflects the deviances we should expect from an observer whose correct responses are binomially distributed with success probability $\psi(x,\hat{\boldsymbol{\theta}})$. A confidence interval for deviance can then be obtained by using the standard percentile method: $D^{*(n)}$ denotes the 100 nth percentile of the distribution **D*** so that, for example, the two-sided 95% confidence interval is written as $[D^{*(.025)}, D^{*(.975)}]$.

Let D_{emp} denote the deviance of our empirically obtained data set. If $D_{\mathrm{emp}} > D^{*(.975)}$, the agreement between data and fit is poor (overdispersion), and it is unlikely that the empirical data set was generated by the best-fitting psychometric function, $\psi(x,\hat{\boldsymbol{\theta}})$. $\psi(x,\hat{\boldsymbol{\theta}})$ is, hence, not an adequate summary of the empirical data or the observer's behavior.

When using Monte Carlo methods to approximate the true deviance distribution **D** by **D***, one requires a large value of B so that the approximation is good enough to be taken as the *true* or *reference* distribution—otherwise, we would simply substitute errors arising from the inappropriate use of an asymptotic distribution for numerical errors incurred by our simulations (Hämmerlin & Hoffmann, 1991).

One way to see whether **D*** has indeed approached **D** is to look at the convergence of several of the quantiles of **D*** with increasing B. For a large range of different values of N, K, and n_i, we found that for $B \geq 10{,}000$, **D*** has stabilized.[14]

Assessing errors in the asymptotic approximation to the deviance distribution. We have found, by a large amount of *trial-and-error* exploration, that errors in the large-sample approximation to the deviance distribution are not predictable in a straightforward manner from one's chosen values of N, K, and **x**.

To illustrate this point, in this section we present six examples in which the χ^2 approximation to the deviance distribution fails in different ways. For each of the six ex-

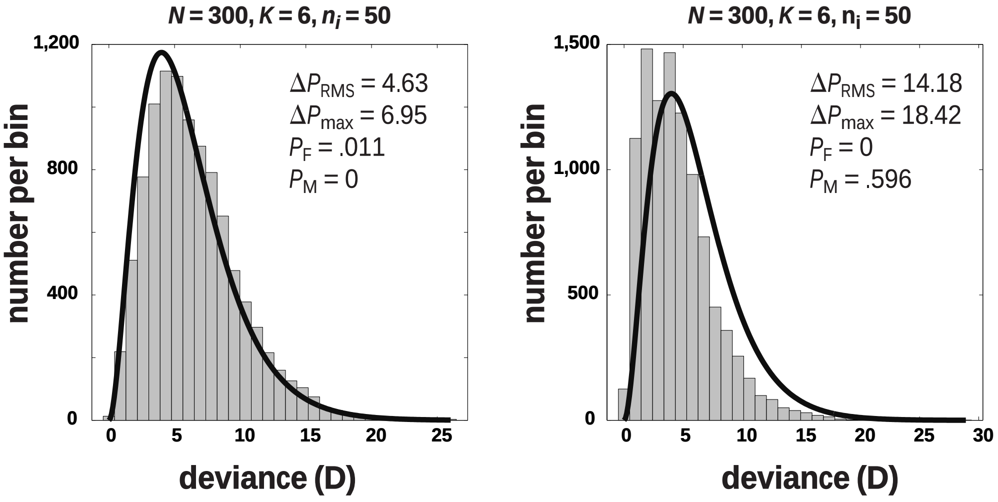

Figure 6. Histograms of Monte-Carlo-generated deviance distributions D* ($B = 10{,}000$). Both panels show distributions for $N = 300$, $K = 6$, and $n_i = 50$. The left-hand panel was generated from $\mathbf{p}_{\mathrm{gen}} = \{.52, .56, .74, .94, .96, .98\}$; the right-hand panel was generated from $\mathbf{p}_{\mathrm{gen}} = \{.63, .82, .89, .97, .99, .9999\}$. The solid dark line drawn with the histograms shows the χ^2_6 distribution (appropriately scaled).

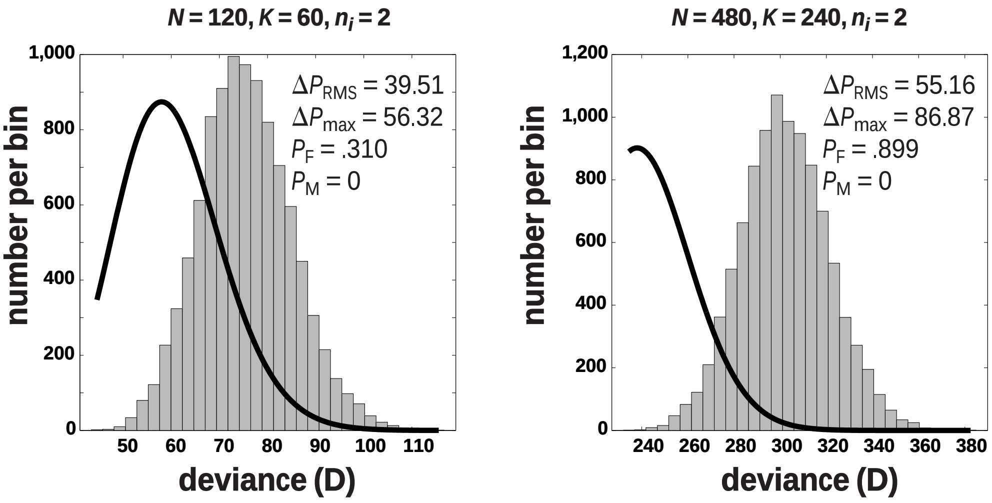

Figure 7. Histograms of Monte-Carlo-generated deviance distributions D* (B = 10,000). For both distributions, $\mathbf{p}_{\text{gen}}$ was uniformly distributed on the interval [.52, .85], and n_i was equal to 2. The left-hand panel was generated using K = 60 (N = 120), and the solid dark line shows χ^2_{60} (appropriately scaled). The right-hand panel's distribution was generated using K = 240 (N = 480), and the solid dark line shows the appropriately scaled χ^2_{240} distribution.

amples, we conducted Monte Carlo simulations, each using $B = 10{,}000$. Critically, for each set of simulations, a set of generating probabilities $\mathbf{p}_{\text{gen}}$ was chosen: In a real experiment, these values would be determined by the positioning of one's sample points **x** and the observer's true psychometric function ψ_{gen}. The specific values of $\mathbf{p}_{\text{gen}}$ in our simulations were chosen by us to demonstrate typical ways in which the χ^2 approximation to the deviance distribution fails: For the two examples shown in Figure 6, we change only $\mathbf{p}_{\text{gen}}$, keeping K and n_i constant; for the two shown in Figure 7, n_i was constant, and $\mathbf{p}_{\text{gen}}$ covered the same range of values; Figure 8, finally, illustrates the effect of changing n_i while keeping $\mathbf{p}_{\text{gen}}$ and K constant.

In order to assess the accuracy of the χ^2 approximation in all examples, we calculated the following four error terms: (1) the rate at which using the χ^2 approximation would have caused us to make Type I errors of rejection, rejecting simulated data sets that should not have been rejected at the 5% level (we call this the *false alarm rate*, or P_F), (2) the rate at which using the χ^2 approximation would have caused us to make Type II errors of rejection, failing to reject a data set that the Monte Carlo distribution **D*** indicated as rejectable at the 5% level (we call this the *miss rate*, P_{M}), (3) the root-mean squared error ΔP_{RMS} in cumulative probability estimate (*CPE*), given by Equation 9 (this is a measure of how different the Monte Carlo distribution **D*** and the appropriate χ^2 distribution are, on average), and (4) the maximal *CPE* error $\Delta P_{\max}$, given by Equation 10, an indication of the maximal error in percentile assignment that could result from using the χ^2 approximation instead of the true **D***.

The first two measures, P_{F} and P_{M}, are primarily useful for individual data sets. The latter two measures, ΔP_{RMS} and $\Delta P_{\max}$, provide useful information in *meta-analyses* (Schmidt, 1996), where models are assessed across several data sets. (In such analyses, we are interested in *CPE* errors even if the deviance value of one particular data set is not close to the tails of **D**: A systematic error in *CPE* in individual data sets might still cause errors of rejection of the model as a whole, when all data sets are considered.)

In order to define ΔP_{RMS} and $\Delta P_{\max}$, it is useful to introduce two additional terms, the Monte Carlo cumulative probability estimate CPE_{MC} and the χ^2 cumulative probability estimate CPE_{χ^2}. By CPE_{MC}, we refer to

$$CPE_{\text{MC}}\left(D\right) = \frac{\#\left\{D_i \le D\right\}}{B+1}, \tag{7}$$

that is, the proportion of deviance values in **D*** smaller than some reference value D of interest. Similarly,

$$CPE_{\chi^2}(D,K) = \int_0^D \chi^2_K\ (x)dx = P\left(\chi^2_K\ \le D\right) \tag{8}$$

provides the same information for the χ^2 distribution with K degrees of freedom. The root-mean squared *CPE* error ΔP_{RMS} (average difference or error) is defined as

$$\Delta P_{RMS} = 100\sqrt{B^{-1}\left(\sum_{i=1}^{B}\left[CPE_{\text{MC}}(D_i) - CPE_{\chi^2}(D_i,K)\right]^2\right)}, \tag{9}$$

and the maximal *CPE* error $\Delta P_{\max}$ (maximal difference or error) is given by

$$\Delta P_{\max} = 100\max\left\{\left|CPE_{\text{MC}}(D_i) - CPE_{\chi^2}\ (D_i,K)\right|\right\}. \tag{10}$$

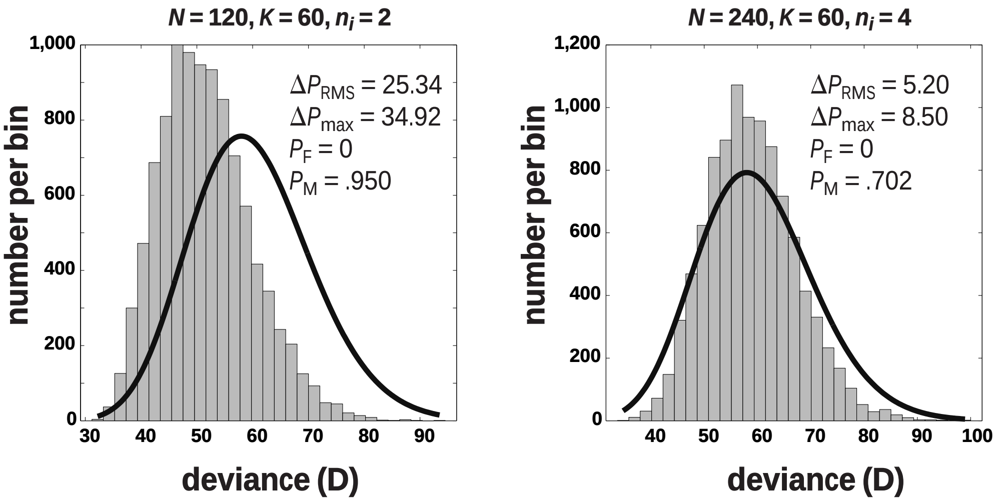

Figure 8. Histograms of Monte-Carlo-generated deviance distributions D* (B = 10,000). For both distributions, $\mathbf{p}_{\text{gen}}$ was uniformly distributed on the interval [.72, .99], and K was equal to 60. The left-hand panel's distribution was generated using $n_i = 2$ ($N = 120$); the right-hand panel's distribution was generated using $n_i = 4$ ($N = 240$). The solid dark line drawn with the histograms shows the χ^2_{60} distribution (appropriately scaled).

Figure 6 illustrates two contrasting ways in which the approximation can fail.[15] The left-hand panel shows results from the test in which the data sets were generated from $\mathbf{p}_{\text{gen}} = \{.52, .56, .74, .94, .96, .98\}$ with $n_i = 50$ observations per sample point ($K = 6$, $N = 300$). Note that the χ^2 approximation to the distribution is (slightly) shifted to the left. This results in $\Delta P_{\text{RMS}} = 4.63$, $\Delta P_{\max} = 6.95$, and a false alarm rate P_{F} of 1.1%. The right-hand panel illustrates the results from very similar input conditions: As before, $n_i = 50$ observations per sample point ($K = 6$, $N = 300$), but now $\mathbf{p}_{\text{gen}} = \{.63, .82, .89, .97, .99, .9999\}$. This time, the χ^2 approximation is shifted to the right, resulting in $\Delta P_{\text{RMS}} = 14.18$, $\Delta P_{\max} = 18.42$, and a large miss rate: $P_{\text{M}} = 59.6\%$.

Note that the reversal is the result of a comparatively subtle change in the distribution of generating probabilities. These two cases illustrate the way in which asymptotic theory may result in errors for sampling schemes that may occur in ordinary experimental settings, using the method of constant stimuli. In our examples, the Type I errors (i.e., erroneously rejecting a valid data set) of the left-hand panel may occur at a low rate, but they do occur. The substantial Type II error rate (i.e., accepting a data set whose deviance is really too high and should thus be rejected) shown on the right-hand panel, however, should be cause for some concern. In any case, the reversal of error type, for the same values of K and n_i, indicates that the type of error is not predictable in any readily apparent way from the distribution of generating probabilities and the error cannot be compensated for by a straightforward correction, such as a manipulation of the number of degrees of freedom of the χ^2 approximation.

It is known that the large-sample approximation of the binomial deviance distribution improves with an increase in n_i (Collett, 1991). In the above examples, n_i was as large as it is likely to get in most psychophysical experiments ($n_i = 50$), but substantial differences between the true deviance distribution and its large-sample χ^2 approximation were nonetheless observed. Increasingly frequently, psychometric functions are fitted to the raw data obtained from adaptive procedures (e.g., Treutwein & Strasburger, 1999), with n_i being considerably smaller. Figure 7 illustrates the profound discrepancy between the true deviance distribution and the χ^2 approximation under these circumstances. For this set of simulations, n_i was equal to 2 for all i. The left-hand panel shows results from the test in which the data sets were generated from $K = 60$ sample points uniformly distributed over [.52, .85] ($\mathbf{p}_{\text{gen}} = \{.52, \ldots, .85\}$), for a total of $N = 120$ observations. This results in $\Delta P_{\text{RMS}} = 39.51$, $\Delta P_{\max} = 56.32$, and a false alarm rate P_{F} of 31.0 %. The right-hand panel illustrates the results from similar input conditions, except that K equaled 240 and N was thus 480. The χ^2 approximation is even worse, with $\Delta P_{\text{RMS}} = 55.16$, $\Delta P_{\max} = 86.87$, and a false alarm rate P_{F} of 89.9%.

The data shown in Figure 7 clearly demonstrate that a large number of observations N, by itself, is not a valid indicator of whether the χ^2 approximation is sufficiently good to be useful for goodness-of-fit assessment.

After showing the effect of a change in $\mathbf{p}_{\text{gen}}$ on the χ^2 approximation in Figure 6, and of K in Figure 7, Figure 8 illustrates the effect of changing n_i while keeping $\mathbf{p}_{\text{gen}}$ and K constant: $K = 60$ and $\mathbf{p}_{\text{gen}}$ were uniformly distributed on the interval [.72, .99]. The left-hand panel shows results from the test in which the data sets were generated with $n_i = 2$ observations per sample point ($K = 60$, $N = 120$). The χ^2 approximation to the distribution is shifted to the right, resulting in $\Delta P_{\text{RMS}} = 25.34$, $\Delta P_{\max} = 34.92$, and a miss rate P_M of 95%. The right-hand panel shows results from very similar generation conditions, except that $n_i = 4$ ($K = 60, N = 240$). Note that, unlike in the other examples introduced so far, the mode of the distribution **D*** is not shifted relative to that of the χ^2_{60} distribution, but that the distribution is more leptokurtic (larger kurtosis or 4th-moment). ΔP_{RMS} equals 5.20, $\Delta P_{\max}$ equals 34.92, and the miss rate P_M is still a substantial 70.2%.

Comparing the left-hand panels of Figures 7 and 8 further points to the impact of **p** on the degree of the χ^2 approximation to the deviance distribution: For constant N, K, and n_i, we obtain either a substantial false alarm rate ($P_F = .31$; Figure 7) or a substantial miss rate ($P_M = .95$; Figure 8).

In general, we have found that very large errors in the χ^2 approximation are relatively rare for $n_i > 40$, but they still remain unpredictable (see Figure 6). For data sets with $n_i < 20$, on the other hand, substantial differences between the true deviance distribution and its large-sample χ^2 approximation are the norm, rather than the exception. We thus feel that Monte-Carlo-based goodness-of-fit assessments should be preferred over χ^2-based methods for binomial deviance.

Deviance residuals. Examination of residuals—the agreement between individual datapoints and the corresponding model prediction—is frequently suggested as being one of the most effective ways of identifying an incorrect model in linear and nonlinear regression (Collett, 1991; Draper & Smith, 1981).

Given that deviance is the appropriate summary statistic, it is sensible to base one's further analyses on the deviance residuals, **d**. Each deviance residual d_i is defined as the square root of the deviance value calculated for datapoint i in isolation, signed according to the direction of the arithmetic residual $y_i - p_i$. For binomial data, this is

$$d_i = \operatorname{sgn}\left(y_i - p_i\right)\sqrt{2\left[n_i y_i \log\left(\frac{y_i}{p_i}\right) + n_i\left(1 - y_i\right)\log\left(\frac{1 - y_i}{1 - p_i}\right)\right]}. \tag{11}$$

Note that

$$D = \sum_{i=1}^{K} d_i^2. \tag{12}$$

Viewed this way, the summary statistic deviance is the sum of the squared deviations between model and data; the d_is are thus analogous to the normally distributed unit-variance deviations that constitute the χ^2 statistic.

Model checking. Over and above inspecting the residuals visually, one simple way of looking at the residuals is to calculate the correlation coefficient between the residuals and the **p** values predicted by one's model. This allows the identification of a systematic (linear) relation between deviance residuals **d** and model predictions **p**, which would suggest that the chosen functional form of the model is inappropriate—for psychometric functions, that presumably means that F is inappropriate.

Needless to say, a correlation coefficient of zero implies neither that there is no systematic relationship between residuals and the model prediction nor that the model chosen is correct; it simply means that whatever relation might exist between residuals and model predictions, it is not a linear one.

Figure 9A shows data from a visual masking experiment with $K = 10$ and $n_i = 50$, together with the best-fitting Weibull psychometric function (Wichmann, 1999). Figure 9B shows a histogram of **D*** for $B = 10{,}000$ with the scaled χ^2_{10}-PDF. The two arrows below the deviance axis mark the two-sided 95% confidence interval [$D^{*(.025)}$, $D^{*(.975)}$]. The deviance of the data set D_{emp} is 8.34, and the Monte Carlo cumulative probability estimate is $CPE_{\text{MC}} = .479$. The summary statistic deviance, hence, does not indicate a lack of fit. Figure 9C shows the deviance residuals **d** as a function of the model prediction **p** [$\mathbf{p} = \psi(x;\hat{\boldsymbol{\theta}})$ in this case, because we are using a fitted psychometric function]. The correlation coefficient between **d** and **p** is $r = -.610$. However, in order to determine whether this correlation coefficient r is significant (of greater magnitude than expected by chance alone if our chosen model was correct), we need to know the expected distribution **r**. For correct models, large samples, and *continuous* data—that is, very large n_i—one should expect the distribution of the correlation coefficients to be a zero-mean Gaussian, but with a variance that itself is a function of **p** and, hence, ultimately of one's sampling scheme **x**. Asymptotic methods are, hence, of very limited applicability for this goodness-of-fit assessment.

Figure 9D shows a histogram of **r***obtained by Monte Carlo simulation with $B = 10{,}000$, again with arrows marking the two-sided 95% confidence interval [$\mathbf{r}^{*(.025)}$, $\mathbf{r}^{*(.975)}$]. Confidence intervals for the correlation coefficient are obtained in a manner analogous to the those obtained for deviance. First, we generate B simulated data sets $\mathbf{y}_i^*$, using the best-fitting psychometric function as generating function. Then, for each synthetic data set $\mathbf{y}_i^*$, we calculate the correlation coefficient r_i^* between the deviance residuals $\mathbf{d}_i^*$ calculated using Equation 11 and the model predictions $\mathbf{p} = \psi(\mathbf{x};\hat{\boldsymbol{\theta}})$. From **r***, one then obtains 95% confidence limits, using the appropriate quantile of the distribution.

In our example, a correlation of $-.610$ is significant, the Monte Carlo cumulative probability estimate being $CPE_{\text{MC}}(-.610) = .0015$. (Note that the distribution is skewed and not centred on zero; a positive correlation of the same magnitude would still be within the 95% confidence interval.) Analyzing the correlation between deviance residuals and model predictions thus allows us to reject the Weibull function as underlying function F for the data shown in Figure 9A, even though the overall deviance does not indicate a lack of fit.

Learning. Analysis of the deviance residuals **d** as a function of temporal order can be used to show perceptual learning, one type of nonstationary observer performance. The approach is equivalent to that described for model checking, except that the correlation coefficient of deviance

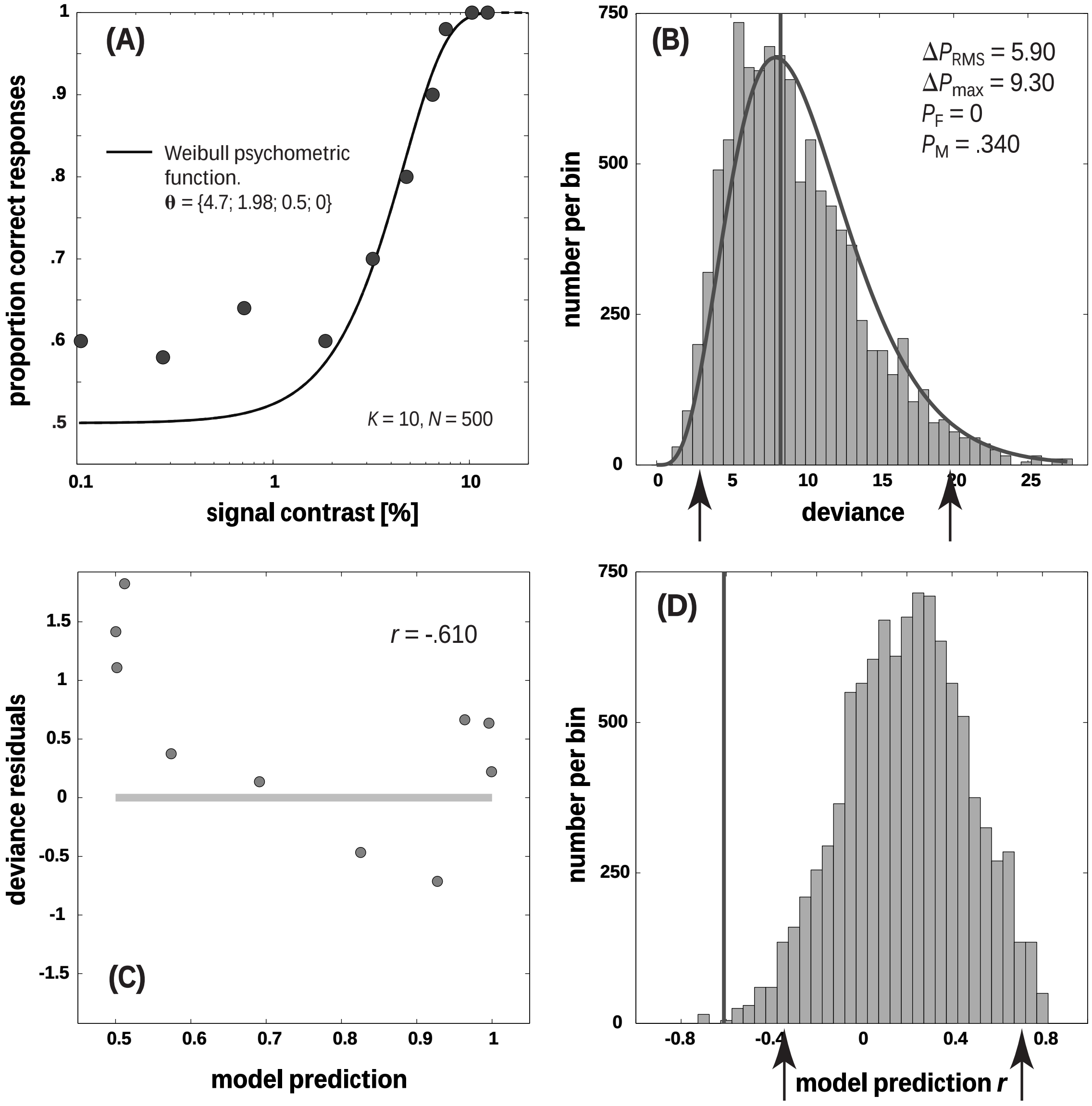

Figure 9. (A) Raw data with $N = 500$, $K = 10$, and $n_i = 50$, together with the best-fitting Weibull psychometric function ψ_{fit} with parameter vector $\theta = \{4.7, 1.98, .5, 0\}$ on semilogarithmic coordinates. (B) Histogram of Monte-Carlo-generated deviance distribution D* ($B = 10{,}000$) from ψ_{fit}. The solid vertical line marks the deviance of the empirical data set shown in panel A, $D_{\text{emp}} = 8.34$; the two arrows below the *x*-axis mark the two-sided 95% confidence interval [$D^{*(.025)}$, $D^{*(.975)}$]. (C) Deviance residuals d plotted as a function of model predictions p on linear coordinates. (D) Histogram of Monte-Carlo-generated correlation coefficients between d and p, r* ($B = 10{,}000$). The solid vertical line marks the correlation coefficient between d and p $= \psi_{\text{fit}}$ of the empirical data set shown in panel A, $\text{r}_{\text{emp}} = -.610$; the two arrows below the *x*-axis mark the two-sided 95% confidence interval [$\text{r}^{*(.025)}$, $\text{r}^{*(.975)}$].

residuals is assessed as a function of the order in which the data were collected (often referred to as their *index*; Collett, 1991). Assuming that perceptual learning improves performance over time, one would expect the fitted psychometric function to be an average of the poor earlier performance and the better later performance.[16] Deviance residuals should thus be negative for the first few datapoints and positive for the last ones. As a consequence, the correlation coefficient *r* of deviance residuals **d** against their indices (which we will denote by **k**) is expected to be positive if the subject's performance improved over time.

Figure 10A shows another data set from one of F.A.W.'s discrimination experiments; again, $K = 10$ and $n_i = 50$, and the best-fitting Weibull psychometric function is shown with the raw data. Figure 10B shows a histogram of **D*** for $B = 10{,}000$ with the scaled χ^2_{10}-PDF. The two arrows below the deviance axis mark the two-sided 95%

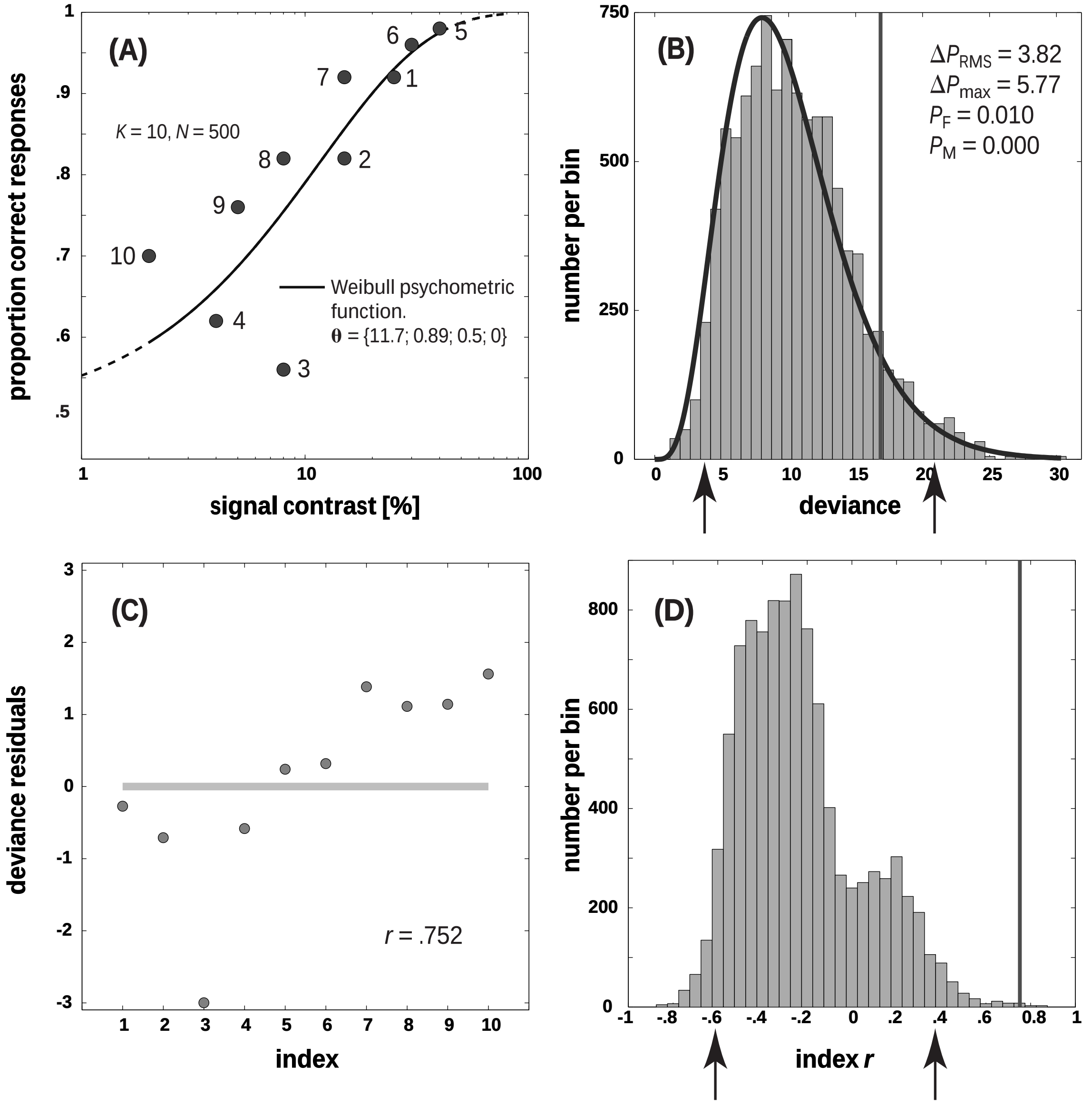

Figure 10. (A) Raw data with $N = 500$, $K = 10$, and $n_i = 50$, together with the best-fitting Weibull psychometric function ψ_{fit} with parameter vector θ = {11.7, 0.89, .5, 0} on semilogarithmic coordinates. The number next to each individual data point shows the index k_i of that data point. (B) Histogram of Monte-Carlo-generated deviance distribution D* ($B = 10{,}000$) from ψ_{fit}. The solid vertical line marks the deviance of the empirical data set shown in panel A, $D_{\text{emp}} = 16.97$; the two arrows below the x-axis mark the two-sided 95% confidence interval [$D^{*(.025)}$, $D^{*(.975)}$]. (C) Deviance residuals d plotted as a function of their index k. (D) Histogram of Monte-Carlo-generated correlation coefficients between d and index k, r* ($B = 10{,}000$). The solid vertical line marks the empirical value of r for the data set shown in panel A, $r_{\text{emp}} = .752$; the two arrows below the x-axis mark the two-sided 95% confidence interval [r*(.025), r*(.975)].

confidence interval [$\mathbf{D}^{*(.025)}$, $\mathbf{D}^{*(.975)}$]. The deviance of the data set D_{emp} is 16.97, the Monte Carlo cumulative probability estimate CPE_{MC}(16.97) being .914. The summary statistic D does not indicate a lack of fit. Figure 10C shows an index plot of the deviance residuals **d**. The correlation coefficient between **d** and **k** is $r = .752$, and the histogram **r*** of shown in Figure 10D indicates that such a high positive correlation is not expected by chance alone. Analysis of the deviance residuals against their index is, hence, an objective means to identify perceptual learning and, thus, reject the fit, even if the summary statistic does not indicate a lack of fit.

Influential observations and outliers. Identification of influential observations and outliers are additional requirements for comprehensive goodness-of-fit assessment.

The jackknife resampling technique. The jackknife is a resampling technique where K data sets, each of size

$K - 1$, are created from the original data set $\mathbf{y}$ by successively omitting one datapoint at a time. The jth jackknife $\mathbf{y}_{(-j)}$ data set is thus the same as $\mathbf{y}$, but with the jth datapoint of $\mathbf{y}$ omitted.[17]

Influential observations. To identify influential observations, we apply the jackknife to the original data set and refit each jackknife data set $\mathbf{y}_{(-j)}$; this yields K parameter vectors $\hat{\boldsymbol{\theta}}_{(-1)}, \ldots, \hat{\boldsymbol{\theta}}_{(-K)}$. Influential observations are those that exert an undue influence on one's inferences—that is, on the estimated parameter vector $\hat{\boldsymbol{\theta}}$—and to this end, we compare $\hat{\boldsymbol{\theta}}_{(-1)}, \ldots, \hat{\boldsymbol{\theta}}_{(-K)}$ with $\hat{\boldsymbol{\theta}}$. If a jackknife parameter set $\hat{\boldsymbol{\theta}}_{(-j)}$ is significantly different from $\hat{\boldsymbol{\theta}}$, the jth datapoint is deemed an influential observation because its inclusion in the data set alters the parameter vector significantly (from $\hat{\boldsymbol{\theta}}_{(-j)}$ to $\hat{\boldsymbol{\theta}}$).

Again, the question arises: What constitutes a significant difference between $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{(-j)}$? No general rules exist to decide at which point $\hat{\boldsymbol{\theta}}_{(-j)}$ and $\hat{\boldsymbol{\theta}}$ are significantly different, but we suggest that one should be wary of one's sampling scheme $\mathbf{x}$ if any one or several of the parameter sets $\hat{\boldsymbol{\theta}}_{(-j)}$ are outside the 95% confidence interval of $\hat{\boldsymbol{\theta}}$. Our companion paper describes a parametric bootstrap method to obtain such confidence intervals for the parameters $\hat{\boldsymbol{\theta}}$. Usually, identifying influential observations implies that more data need to be collected, at or around the influential datapoint(s).

Outliers. Like the test for influential observations, this objective procedure to detect outliers employs the jackknife resampling technique. (Testing for outliers is sometimes referred to as *test of discordancy*; Collett, 1991.) The test is based on a desirable property of deviance—namely, its nestedness: Deviance can be used to compare different models for binomial data as long as they are members of the same family. Suppose a model M_1 is a special case of model M_2 (M_1 is "nested within" M_2), so M_1 has fewer free parameters than M_2. We denote the degrees of freedom of the models by v_1 and v_2, respectively. Let the deviance of model M_1 be D_1 and of M_2 be D_2. Then, the difference in deviance, $D_1 - D_2$, has an approximate χ^2 distribution with $v_1 - v_2$ degrees of freedom. This approximation to the χ^2 distribution is usually very good even if each individual distribution, D_1 or D_2, is not reliably approximated by a χ^2 distribution (Collett, 1991); indeed, $D_1 - D_2$, has an approximate χ^2 distribution with $v_1 - v_2$ degrees of freedom, even for binary data, despite the fact that, for binary data, deviance is not even asymptotically distributed according to χ^2 (Collett, 1991). This property makes this particular test of discordancy applicable to (small-sample) psychophysical data sets.

To test for outliers, we again denote the original data set by $\mathbf{y}$ and its deviance by D. In the terminology of the preceding paragraph, the fitted psychometric function $\psi(x;\hat{\boldsymbol{\theta}})$ corresponds to model M_1. Then, the jackknife is applied to $\mathbf{y}$, and each jackknife data set $\mathbf{y}_{(-j)}$ is refit to give K parameter vectors $\hat{\boldsymbol{\theta}}_{(-1)}, \ldots, \hat{\boldsymbol{\theta}}_{(-K)}$, from which to calculate deviance, yielding $D_{(-1)}, \ldots, D_{(-K)}$. For each of the K jackknife parameter vectors $\hat{\boldsymbol{\theta}}_{(-j)}$, an alternative model M_2 for the (complete) original data set $\mathbf{y}$ is constructed as

$$M_2 : \begin{cases} p_i = \psi(x, \hat{\boldsymbol{\theta}}_{(-j)}) & \text{if } x_i \neq x_j \\ p_i = \psi(x, \hat{\boldsymbol{\theta}}_{(-j)}) + \zeta & \text{if } x_i = x_j \end{cases}. \quad (13)$$

Setting ζ equal to $y_j - \psi(x;\hat{\boldsymbol{\theta}}_{(-j)})$, the deviance of M_2 equals $D_{(-j)}$, because the jth datapoint, dropped during the jackknife, is perfectly fit by M_2 owing to the addition of a dedicated free parameter, ζ.

To decide whether the reduction in deviance, $D - D_{(-j)}$, is significant, we compare it against the χ^2 distribution with one degree of freedom, because $v_1 - v_2 = 1$. Choosing a one-sided 99% confidence interval, M_2 is a better model than M_1 if $D - D_{(-j)} > 6.63$, because $CPE_{\chi^2}(6.63) = .99$. Obtaining a significant reduction in deviance for data set $\mathbf{y}_{(-j)}$ implies that the jth datapoint is so far away from the original fit $\psi(x;\hat{\boldsymbol{\theta}})$ that the addition of a dedicated parameter ζ, whose sole function is to fit the jth datapoint, reduces overall deviance significantly. Datapoint j is thus very likely an outlier, and as in the case of influential observations, the best strategy generally is to gather additional data at stimulus intensity x_j, before more radical steps, such as removal of y_j from one's data set, are considered.

Discussion

In the preceding sections, we introduced statistical tests to identify the following: first, inappropriate choice of F; second, perceptual learning; third, an objective test to identify influential observations; and finally, an objective test to identify outliers. The histograms shown in Figures 9D and 10D show the respective distributions $\mathbf{r}^*$ to be skewed and not centered on zero. Unlike our summary statistic D, where a large-sample approximation for binomial data with $n_i > 1$ exists even if its applicability is sometimes limited, neither of the correlation coefficient statistics has a distribution for which even a roughly correct asymptotic approximation can easily be found for the K, N, and $\mathbf{x}$ typically used in psychophysical experiments. Monte Carlo methods are thus without substitute for these statistics.

Figures 9 and 10 also provide another good demonstration of our warnings concerning the χ^2 approximation of deviance. It is interesting to note that for both data sets, the MCS deviance histograms shown in Figures 9B and 10B, when compared against the asymptotic χ^2 distributions, have considerable ΔP_{RMS} values of 3.8 and 5.9, with $\Delta P_{\max} = 5.8$ and 9.3, respectively. Furthermore, the miss rate in Figure 9B is very high ($P_{\mathrm{M}} = .34$). This is despite a comparatively large number of trials in total and per block ($N = 500$, $n_i = 50$), for both data sets. Furthermore, whereas the χ^2 is shifted toward higher deviances in Figure 9B, it is shifted toward lower deviance values in Figure 10B. This again illustrates the complex interaction between deviance and $\mathbf{p}$.

SUMMARY AND CONCLUSIONS

In this paper, we have given an account of the procedures we use to estimate the parameters of psychometric functions and derive estimates of thresholds and slopes. An essential part of the fitting procedure is an assessment of goodness of fit, in order to validate our estimates.

We have described a constrained maximum-likelihood algorithm for fitting three-parameter psychometric functions to psychophysical data. The third parameter, which specifies the upper asymptote of the curve, is highly constrained, but it can be shown to be essential for avoiding bias in cases where observers make stimulus-independent errors, or lapses. In our laboratory, we have found that the lapse rate for trained observers is typically between 0% and 5%, which is enough to bias parameter estimates significantly.

We have also described several goodness-of-fit statistics, all of which rely on resampling techniques to generate accurate approximations to their respective distribution functions or to test for influential observations and outliers. Fortunately, the recent sharp increase in computer processing speeds has made it possible to fulfill this computationally expensive demand. Assessing goodness of fit is necessary in order to ensure that our estimates of thresholds and slopes, and their variability, are generated from a plausible model for the data and to identify problems with the data themselves, be they due to learning, to uneven sampling (resulting in influential observations), or to outliers.

Together with our companion paper (Wichmann & Hill, 2001), we cover the three central aspects of modeling experimental data: parameter estimation, obtaining error estimates on these parameters, and assessing goodness of fit between model and data.

REFERENCES

Collett, D. (1991). *Modeling binary data*. New York: Chapman & Hall/CRC.

Dobson, A. J. (1990). *Introduction to generalized linear models*. London: Chapman & Hall.

Draper, N. R., & Smith, H. (1981). *Applied regression analysis*. New York: Wiley.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, **7**, 1-26.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans* (CBMS-NSF Regional Conference Series in Applied Mathematics). Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician*, **37**, 36-48.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Finney, D. J. (1952). *Probit analysis* (2nd ed.). Cambridge: Cambridge University Press.

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

Forster, M. R. (1999). Model selection in science: The problem of language variance. *British Journal for the Philosophy of Science*, **50**, 83-102.

Gelman, A. B., Carlin, J. S., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. New York: Chapman & Hall/CRC.

Hämmerlin, G., & Hoffmann, K.-H. (1991). *Numerical mathematics* (L. T. Schumacher, Trans.). New York: Springer-Verlag.

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

Hinkley, D. V. (1988). Bootstrap methods. *Journal of the Royal Statistical Society B*, **50**, 321-337.

Hoel, P. G. (1984). *Introduction to mathematical statistics*. New York: Wiley.

Lam, C. F., Mills, J. H., & Dubno, J. R. (1996). Placement of observations for the efficient estimation of a psychometric function. *Journal of the Acoustical Society of America*, **99**, 3689-3693.

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

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. London: Chapman & Hall.

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.

O'Regan, J. K., & Humbert, R. (1989). Estimating psychometric functions in forced-choice situations: Significant biases found in threshold and slope estimation 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.). New York: Cambridge University Press.

Quick, R. F. (1974). A vector magnitude model of contrast detection. *Kybernetik*, **16**, 65-67.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, **1**, 115-129.

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

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

Weibull, W. (1951). Statistical distribution function of wide applicability. *Journal of Applied Mechanics*, **18**, 292-297.

Wichmann, F. A. (1999). *Some aspects of modelling human spatial vision: Contrast discrimination*. Unpublished doctoral dissertation, Oxford University.

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

NOTES

1. For illustrative purposes, we shall use the Weibull function for F (Quick, 1974; Weibull, 1951). This choice was based on the fact that the Weibull function generally provides a good model for contrast discrimination and detection data (Nachmias, 1981) of the type collected by one of us (F.A.W.) over the past few years. It is described by

$$F(x;\alpha;\beta) = 1 - \exp\left[-\left(\frac{x}{\alpha}\right)^{\beta}\right], \qquad 0 \le x < \infty.$$

2. Often, particularly in studies using forced-choice paradigms, λ does not appear in the equation, because it is fixed at zero. We shall illustrate and investigate the potential dangers of doing this.

3. The simplex search method is reliable but converges somewhat slowly. We choose to use it for ease of implementation: first, because of its reliability in approximating the global minimum of an error surface, given a good initial guess, and second, because it does not rely on gradient descent and is therefore not catastrophically affected by the sharp increases in the error surface introduced by our Bayesian priors (see the next section). We have found that the limitations on its precision (given

error tolerances that allow the algorithm to complete in a reasonable amount of time on modern computers) are many orders of magnitude smaller than the confidence intervals estimated by the bootstrap procedure, given psychophysical data and are therefore immaterial for the purposes of fitting psychometric functions.

4. In terms of Bayesian terminology, our prior $W(\lambda)$ is not a *proper* prior density, because it does not integrate to 1 (Gelman, Carlin, Stern, & Rubin, 1995). However, it integrates to a positive finite value that is reflected, in the log-likelihood surface, as a constant offset that does not affect the estimation process. Such prior densities are generally referred to as *unnormalized densities*, distinct from the sometimes problematic *improper* priors that do not integrate to a finite value.

5. See Treutwein and Strasburger (1999) for a discussion of the use of beta functions as Bayesian priors in psychometric function fitting. Flat priors are frequently referred to as (maximally) *noninformative* priors in the context of Bayesian data analysis, to stress the fact that they ensure that inferences are unaffected by information external to the current data set (Gelman et al., 1995).

6. One's choice of prior should respect the implementation of the search algorithm used in fitting. Using the flat prior in the above example, an increase in λ from .06 to .0600001 causes the maximization term to jump from zero to negative infinity. This would be catastrophic for some gradient-descent search algorithms. The simplex algorithm, on the other hand, simply withdraws the step that took it into the "infinitely unlikely" region of parameter space and continues in another direction.

7. The log-likelihood error metric is also extremely sensitive to very *low* predicted performance values (close to 0). This means that, in yes/no paradigms, the same arguments will apply to assumptions about the lower bound as those we discuss here in the context of λ. In our 2AFC examples, however, the problem never arises, because γ is fixed at .5.

8. In fact, there is another reason why λ needs to be tightly constrained: It covaries with α and β, and we need to minimize its negative impact on the estimation precision of α and β. This issue is taken up in our Discussion and Summary section.

9. In this case, the noise scheme corresponds to what Swanson and Birch (1992) call "extraneous noise." They showed that extraneous noise can bias threshold estimates, in both the method of constant stimuli and adaptive procedures with the small numbers of trails commonly used within clinical settings or when testing infants. We have also run simulations to investigate an alternative noise scheme, in which λ_{gen} varies between blocks in the same dataset: A new λ_{gen} was chosen for each block from a uniform random distribution on the interval [0, .05]. The results (not shown) were not noticeably different, when plotted in the format of Figure 3, from the results for a fixed λ_{gen} of .2 or .3.

10. Uniform random variates were generated on the interval (0,1), using the procedure **ran2 ()** from Press et al. (1992).

11. This is a crude approximation only; the actual value depends heavily on the sampling scheme. See our companion paper (Wichmann & Hill, 2001) for a detailed analysis of these dependencies.

12. In rare cases, underdispersion may be a direct result of observers' behavior. This can occur if there is a negative correlation between individual binary responses and the order in which they occur (Colett, 1991). Another hypothetical case occurs when observers use different cues to solve a task and switch between them on a nonrandom basis during a block of trials (see the Appendix for proof).

13. Our convention is to compare deviance, which reflects the probability of obtaining $\mathbf{y}$ given $\hat{\boldsymbol{\theta}}$, against a distribution of probability measures of $\mathbf{y}^*_1 \ldots \mathbf{y}^*_B$, each of which is *also* calculated assuming $\hat{\boldsymbol{\theta}}$. Thus, the test assesses whether the data are *consistent* with having been generated by our fitted psychometric function; it does not take into account the number of free parameters in the psychometric function used to obtain $\hat{\boldsymbol{\theta}}$. In these circumstances, we can expect, for suitably large data sets, D to be distributed as χ^2 with K degrees of freedom. An alternative would be to use the maximum-likelihood parameter estimate for *each* simulated data set, so that our simulated deviance values reflect the probabilities of obtaining $\mathbf{y}^*_1 \ldots \mathbf{y}^*_{\text{B}}$ given $\hat{\boldsymbol{\theta}}^*_1 \ldots \hat{\boldsymbol{\theta}}^*_B$. Under the latter circumstances, the expected distribution has $K - P$ degrees of freedom, where P is the number of parameters of the discrepancy function (which is often, but not always, well approximated by the number of free parameters in one's model—see Forster, 1999). This procedure is appropriate if we are interested not merely in fitting the data (summarizing, or replacing, data by a fitted function), but in *modeling data*, or *model comparison*, where the particulars of the model(s) itself are of interest.

14. One of several ways we assessed convergence was to look at the quantiles .01, .05, .1, .16, .25, .5, .75, .84, .9, .95, and .99 of the simulated distributions and to calculate the root mean square (RMS) percentage change in these deviance values as B increased. An increase from B = 500 to B = 500,000, for example, resulted in an RMS change of approximately 2.8%, whereas an increase from B = 10,000 to B = 500,000 gave only 0.25%, indicating that for B = 10,000, the distribution has already stabilized. Very similar results were obtained for all sampling schemes.

15. The differences are even larger if one does not exclude datapoints for which model predictions are $p = 0$ or $p = 1.0$, because such points have zero deviance (zero variance). Without exclusion of such points, χ^2-based assessment systematically overestimates goodness of fit. Our Monte Carlo goodness-of-fit method, on the other hand, is accurate whether such points are removed or not.

16. For this statistic, it is important to remove points with $y = 1.0$ or $y = 0.0$ to avoid errors in one's analysis.

17. Jackknife data sets have negative indices inside the brackets as a reminder that the jth datapoint has been removed from the original data set in order to create the jth jackknife data set. Note the important distinction between the more usual connotation of "jackknife," in which single observations are removed sequentially, and our coarser method, which involves removal of whole blocks at a time. The fact that observations in different blocks are not identically distributed and that their generating probabilities are parametrically related by $\psi(x,\hat{\boldsymbol{\theta}})$ may make our version of the jackknife unsuitable for many of the purposes (such as variance estimation) to which the conventional jackknife is applied (Efron, 1979, 1982; Efron & Gong, 1983; Efron & Tibshirani, 1993).

APPENDIX
Variance of Switching Observer

Assume an observer with two cues, c_1 and c_2, at his or her disposal, with associated success probabilities p_1 and p_2, respectively. Given N trials, the observer chooses to use cue c_1 on Nq of the trials and c_2 on $N(1-q)$ of the trials. Note that q is not a probability, but a fixed fraction: The observer uses c_1 always and on exactly Nq of the trials. The expected number of correct responses of such an observer is

$$E_s = N\left[qp_1 + \left(1-q\right)p_2\right]. \tag{A1}$$

The variance of the responses around E_s is given by

$$\sigma_s^2 = N\left[qp_1\left(1-p_1\right) + \left(1-q\right)p_2\left(1-p_2\right)\right]. \tag{A2}$$

Binomial variance, on the other hand, with $E_b = E_s$ and, hence, $p_b = qp_1 + (1-q)p_2$, equals

$$\sigma_b^2 = Np_b\left(1-p_b\right) = N\left(qp_1 + \left(1-q\right)p_2\right)\left[1-\left(qp_1+\left(1-q\right)p_2\right)\right]. \tag{A3}$$

For $q = 0$ or $q = 1$ Equations A2 and A3 reduce to $\sigma_s^2 = \sigma_b^2 = Np_2(1-p_2)$ and $\sigma_s^2 = \sigma_b^2 = Np_1(1-p_1)$, respectively. However, simple algebraic manipulation shows that for $0 < q < 1$, $\sigma_s^2 < \sigma_b^2$ for all $p_1, p_2 \in [0,1]$ if $p_1 \neq p_2$.

Thus, the variance of such a "fixed-proportion-switching" observer is smaller than that of a binomial distribution with the same expected number of correct responses. This is an example of underdispersion that is inherent in the observer's behavior.

(Manuscript received June 10, 1998;
revision accepted for publication February 27, 2001.)