Psychological Bulletin
1991, Vol. 109, No. 1, 152-159

0033-2909/91/$3.00

Thresholds From Psychometric Functions: Superiority of Bootstrap to Incremental and Probit Variance Estimators

David H. Foster
Department of Communication and Neuroscience
University of Keele, Staffordshire, England

Walter F. Bischof
Alberta Centre for Machine Intelligence and Robotics
University of Edmonton, Alberta, Edmonton, Canada

The bootstrap method provides a powerful, general procedure for estimating the variance of a parameter of a function. The parametric version of the method was used to estimate the standard deviation of a threshold from a psychometric function and the standard deviation of its slope. Bootstrap standard deviations were compared with those obtained by a classical incremental method and by the asymptotic method of probit analysis. Twelve representative experimental conditions were tested in Monte Carlo studies, each of 1,000 data sets. All methods performed equally well with large data sets, but with small data sets the bootstrap was superior in both percentage bias and relative efficiency.

There are many occasions in which it is desirable to measure the strength of a stimulus in terms of its response in an organism. Typically, different levels of a known treatment are applied to subjects and the effects of that treatment are recorded at each level. Thus, in psychophysics, one might construct a *psychometric function*, which describes the relationship between the level of a stimulus and the probability of a subject making a particular response at that level (Falmagne, 1982). In a biological or medical assay, one might determine a *stimulus–response curve* or *dose–response curve*, which relates the dosage of a drug or poison and the proportion of subjects that on average are affected at that dosage (Finney, 1978).

In practice, the potency of a stimulus may need to be characterized by a single number that corresponds to a particular criterion level of efficacy. For a psychometric function, this stimulus level is the *threshold* value of the stimulus, for that particular criterion. In a simple "yes–no" detection task, percentage of successes might be recorded at a number of testing levels and a theoretical function in the form, for example, of a normal probability integral function fitted to those data. The situation is illustrated in Figure 1a. Threshold would be defined for a criterion performance level of 50%. For a two-alternative forced-choice task, where theoretical performance ranges from 50% to 100%, the criterion level could be 75%. For a dose–response curve, the situation is similar. The criterion level of efficacy would be the *median* (or *mean*) *effective dose*, symbolized by ED50, which on average produces a response in 50% of subjects. Similarly, ED75 is the dose that produces a response in 75% of subjects.

How stimulus levels should be sampled to best obtain a threshold estimate has been the subject of some attention in the literature (see, e.g., Emerson, 1984; Shelton, Picardi, & Green, 1982; Taylor, Forbes, & Creelman, 1983, for reviews of adaptive and other methods in psychophysics; see Finney, 1978, for discussion of methods relevant to medical and biological assay). Less attention has been directed to the problem of estimating the *reliability* of a threshold or a median-effective-dose estimate. In some circumstances, the question may be resolved empirically: The experiment is repeated a number of times and the precision of an individual estimate or mean of estimates is estimated from the sample variance. In other circumstances, repeating the experiment may be impossible or impracticable. It may still be important, however, to obtain information about the reliability of a single estimate, for example, when judging whether the estimate is significantly different from another obtained from a different subject or under different experimental conditions. The question has particular significance in assay work when deciding on the minimum number of subjects from which an acceptably precise ED50 may be calculated.

Probit analysis has been the traditional method for estimating the variance or standard deviation of a threshold estimate from a psychometric function (Finney, 1952, 1971). The binomial scores at each testing level are transformed (by the inverse of the normal probability integral), a straight line is fitted by a weighted linear regression, and a threshold (ED50) computed. The probit method has been very popular. There have been over 2,300 citations of Finney's *Probit Analysis* (1947, 1952, and 1971 editions) over the 10-year period from 1978 to 1988. In the method, the standard deviation of the threshold estimate is obtained by classical asymptotic theory. The trustworthiness of the estimate, however, is uncertain when sample sizes are not large (Finney, 1952, pp. 250–251; 1971, p. 57), and examples of substantial errors have been reported (Foster & Bischof, 1987; McKee, Klein, & Teller, 1985).

The bootstrap procedure (Efron, 1982; Efron & Tibshirani, 1986) for estimating the standard deviation of a point estimate

We thank P. Jones for advice and B. Efron and D. J. Finney for critical review of the manuscript of this article. Walter F. Bischof was assisted by Grant No. 81.166.0.84 from the Swiss National Science Foundation.

Correspondence concerning this article should be addressed to David H. Foster, Department of Communication and Neuroscience, University of Keele, Keele, Staffordshire ST5 5BG England.

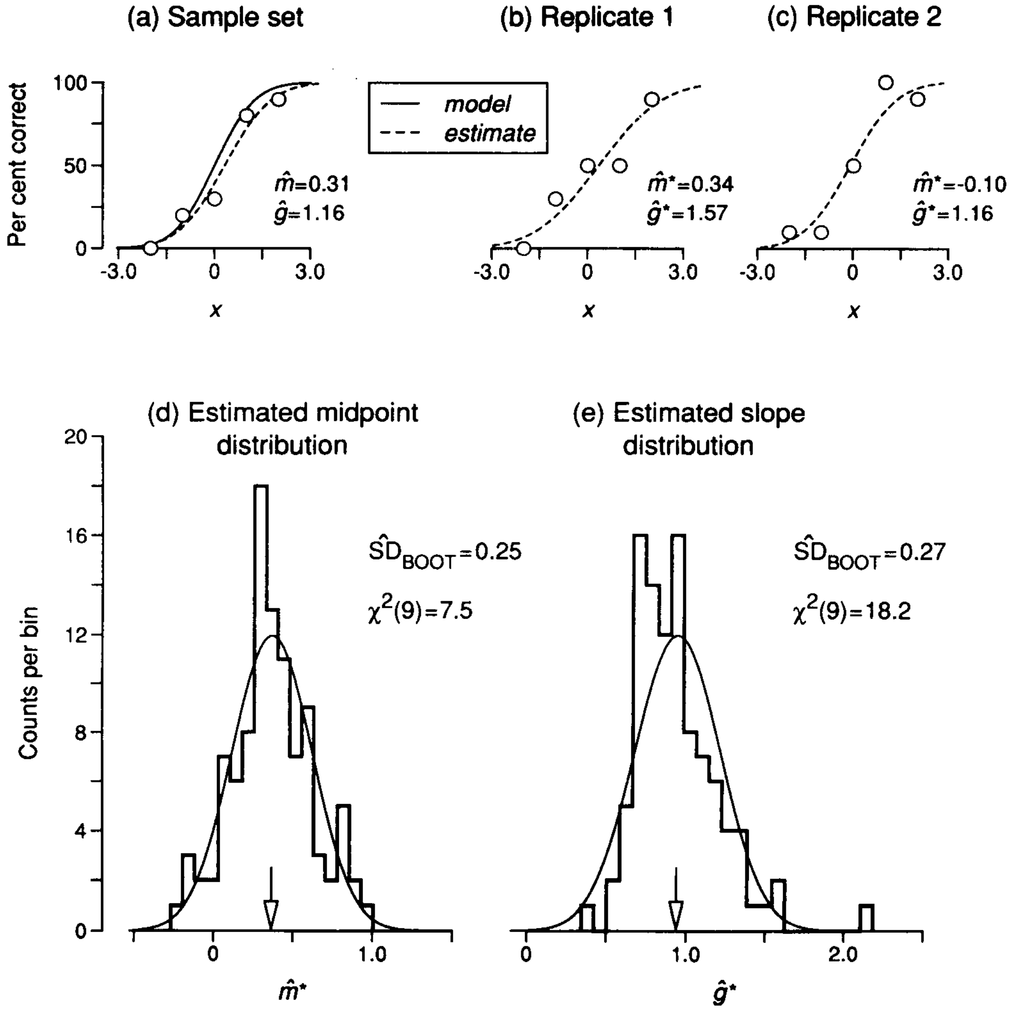

Figure 1. Bootstrap standard deviations from a psychometric function: (a) Model psychometric function (continuous curve) with a sample set of data (open symbols), based on Equation 1 with $m = 0, g = 1, n_i = n = 10$, and maximum likelihood estimates of the psychometric function (broken curve) and of the midpoint $\hat{m}$ and slope $\hat{g}$; (b) and (c) are bootstrap replicates generated from the sample set in (a); (d) and (e) are histograms of the 100 bootstrap replicates $\hat{m}^*$ and $\hat{g}^*$. (The smooth curves are normal distributions with the same means and standard deviations as the bootstrap histograms. Goodness of fit is shown by the chi-squared values.)

(or any other aspect of a distribution) is essentially a Monte Carlo sampling technique. The following is an example (Efron & Gong, 1983) used to illustrate the bootstrap. Consider 15 pairs of average test scores from 15 law schools, each pair of scores comprising two different measures of subject performance (the average undergraduate grade-point average and the average score on the law school admission test). The observed Pearson correlation coefficient r for these 15 pairs was .776. The bootstrap estimate of the precision of this estimate is obtained as follows. The original 15 pairs are each copied a very large number of times (say one billion) and mixed together. Samples of size 15 are then selected at random and values of r calculated for each sample. A typical bootstrap sample might consist of 2 copies of the first pair of original values, 0 copies of the second pair, 1 copy of the third pair, and so forth, the total number summing to 15. This process is repeated a large number of times, say 1,000, to obtain 1,000 bootstrap estimates of r. The standard deviation of these 1,000 estimates constitutes the bootstrap estimate of the standard deviation. For the law-school data, the bootstrap standard deviation was .127, which may be compared with the normal theory estimate of .115 (Efron & Gong, 1983).

The application of the bootstrap procedure in the present context is similar. A large number of samples is randomly drawn, with replacement, from the original set of data values giving response as a function of stimulus level. (This sampling process may be improved by using "smoothed" versions of the original data estimated from the fitted psychometric function.) Each of these bootstrap samples is fitted by the psychometric function and a threshold estimate calculated. The standard deviation of the resulting distribution of bootstrap estimates of the threshold is used to estimate the standard deviation of the threshold obtained from the original data set.

As Efron (1982) emphasized, the success of the bootstrap method depends on replacing traditional theoretical analysis by computing effort. It requires few modeling assumptions and little theoretical analysis. One of its advantages in the present context is its potential small-sample accuracy (Hinkley, 1988).

The purpose of the present study was to compare the probit and bootstrap methods and a third, incremental method (Foster, 1986) based on the use of a Taylor-series expansion of the threshold estimate as a function of the empirical data. The variables of interest were the standard deviation of the midpoint of the estimated psychometric function (corresponding to the threshold test level) and the standard deviation of the slope of the estimated psychometric function at its midpoint. A representative range of experimental conditions was defined, with different spacings of the test levels and different numbers of trials per level. For each experimental condition, 1,000 Monte Carlo sets of data were generated, to which each of the three methods for estimating the standard deviation was applied. The performance of each of the methods was judged by two statistics: the *percentage bias* of the standard deviation estimator and the *relative efficiency* of the standard deviation estimator. The bootstrap method was found to be superior to the probit and incremental methods, particularly in the analysis of small data sets.

Method

Let $Y_1, Y_2, \ldots, Y_l$ be an observed set of scores measured at l test levels, $x_1, x_2, \ldots, x_l$, of the stimulus. Each score Y_i represents the proportion of r_i successes out of n_i trials, $Y_i = r_i/n_i$. The underlying psychometric function is assumed to have the form of the normal probability integral

$$y = \Phi(z) = (2\pi)^{-1/2} \int_{-\infty}^{z} \exp(-u^2/2)du, \tag{1a}$$

$$z = (x - m)\cdot g, \tag{1b}$$

where the constants m and g define the midpoint of the function and the gradient or slope at the midpoint (except for the factor $(2\pi)^{-1/2}$). The symbol g should not be confused with the symbol defined by Finney (1971, p. 78) for another purpose. The observed scores Y_i are assumed to be generated from rescaled binomial distributions

$$Y_i \sim \mathrm{Bi}(n_i, y_i)/n_i, \; i = 1, 2, \ldots, l, \tag{1c}$$

where $y_i = y$ at $x = x_i$. This analysis is not especially dependent on the choice of the normal probability integral function, and other functions, such as the logistic function, would be acceptable; see Finney (1971) and Cox (1970).

In Figure 1a, the continuous curve shows an example of the model function of Equations 1a and 1b, with $m = 0$, $g = 1$, the open symbols a sample set generated from Equation 1c with $n_i = 10$, $l = 5$, the broken curve the maximum likelihood estimate of Equations 1a and 1b, and $\hat{m}$ and $\hat{g}$ the maximum likelihood estimates of m and g respectively.

Estimation of Standard Deviation by Probit Method

In the original probit method, maximum likelihood estimates $\hat{m}$ and $\hat{g}$ of m and g are obtained from the observed scores Y_i, $i = 1, 2, \ldots, l$, by an iterative procedure (the Newton-Raphson method). Let $Z_i = z_i + (Y_i - y_i)/(\partial y_i/\partial z_i)$, where $z_i = \Phi^{-1}(Y_i)$ and Φ^{-1} is the inverse of the normal probability integral Φ of Equation 1a. In each iteration a weighted linear regression of Z_i on x_i is computed, with weights $n_i w_i$, $w_i = (\partial y_i/\partial z_i)^2/(y_i(1 - y_i))$. The estimated variances $\hat{\mathrm{V}}$ are given by the asymptotic formulae

$$\hat{\mathrm{V}}(\hat{m}) = -1/(\partial^2 L/\partial \hat{m}^2), \tag{2a}$$

$$\hat{\mathrm{V}}(\hat{g}) = -1/(\partial^2 L/\partial \hat{g}^2), \tag{2b}$$

where L is the likelihood. Equations 2a and 2b lead (Finney, 1952, 1971) to the following computational expressions for the estimated standard deviations $\widehat{\mathrm{SD}}_{\mathrm{PROB}}$

$$\widehat{\mathrm{SD}}_{\mathrm{PROB}}(\hat{m}) = [(1/\textstyle\sum_i n_i w_i + (\hat{m} - x.)^2/\sum_i n_i w_i (x_i - x.)^2)/\hat{g}^2]^{1/2},$$

$$\widehat{\mathrm{SD}}_{\mathrm{PROB}}(\hat{g}) = [1/\textstyle\sum_i n_i w_i (x_i - x.)^2]^{1/2},$$

where $x. = \sum_i n_i w_i x_i/\sum_i n_i w_i$. Details are given in Finney (1952, 1971). Note that Equations 2a and 2b relate properly to large samples. This iterative weighted regression is not essential to the probit method, and Finney (1971, sections 5.4, 6.6) has advocated a direct approach to the maximization of the likelihood. The asymptotic formulae (Equations 2a and 2b) remain unaltered. The principle of the probit method itself may be traced back to Fechner (1860).[1]

Estimation of Standard Deviation by Incremental Method

Consider the functions f_m and f_g defined by $\hat{m} = f_m(Y_1, Y_2, \ldots, Y_l)$ and $\hat{g} = f_g(Y_1, Y_2, \ldots, Y_l)$. Suppose that the estimated variances $\hat{\sigma}_i{}^2$ of the Y_i are not too large (see Lindley, 1965). The estimated standard deviations $\widehat{\mathrm{SD}}_{\mathrm{INC}}$ are then given approximately by the first terms of a Taylor series expansion

$$\widehat{\mathrm{SD}}_{\mathrm{INC}}(\hat{m}) = [\textstyle\sum_i (\partial f_m/\partial Y_i)^2\, \hat{\sigma}_i{}^2]^{1/2}, \tag{3a}$$

$$\widehat{\mathrm{SD}}_{\mathrm{INC}}(\hat{g}) = [\textstyle\sum_i (\partial f_g/\partial Y_i)^2\, \hat{\sigma}_i{}^2]^{1/2}, \tag{3b}$$

where the partial derivatives $\partial f_m/\partial Y_i$, $\partial f_g/\partial Y_i$ are evaluated at $(Y_1, Y_2, \ldots, Y_l)$ (Foster, 1986). The $\hat{\sigma}_i^2$ are given by the usual binomial formula $Y_i(1 - Y_i)/n_i$. This method belongs to the classical study of the "Combination of Observations" (Lindley, 1965). To avoid spuriously small standard deviation estimates from sample sets in which several of the Y_i were 0 or 1, the actual sample data values $Y_1, Y_2, \ldots, Y_l$ were smoothed by replacing each Y_i by $\hat{y}_i$ estimated from the fitted curve of Equations 1a and 1b.[2] This is the *parametric* version of the incremental method (Efron, 1982).

Estimation of Standard Deviation by Bootstrap Method

Consider the empirical distribution of $(Y_1, Y_2, \ldots, Y_l)$, that is the distribution obtained by placing the rescaled binomial $\mathrm{Bi}(n_i, Y_i)/n_i$ at each level x_i, $i = 1, 2, \ldots, l$, of the empirical data set. As in the incremental method, the parametric version of the bootstrap method (Efron, 1982) was used to avoid the effects of several of the Y_i being 0 or 1. Thus the actual sample data values Y_i in $\mathrm{Bi}(n_i, Y_i)/n_i$ were replaced

[1] See Fechner (1860), Chapter 8, Section 1d, "Specielles zur Methode der richtigen und falschen Faelle" (special [comments] on the method of correct and incorrect cases). Fechner's calculations were verified by Moebius.

[2] When $Y_i = 0$ or 1, the estimated standard deviations $\hat{\sigma}_i{}^2 = 0$ and contribute nothing to the estimates $\widehat{\mathrm{SD}}(\hat{m})$ and $\widehat{\mathrm{SD}}(\hat{g})$.

Table 1
Comparison of Bootstrap, Incremental, and Probit Estimators for the Standard Deviation of the Estimated Midpoint ($\hat{m}$) and Gradient ($\hat{g}$) for Model Function (Equation 1)

		Bootstrap estimate ($\widehat{\mathrm{SD}}_{\mathrm{BOOT}}$)				Incremental estimate ($\widehat{\mathrm{SD}}_{\mathrm{INC}}$)				Probit estimate ($\widehat{\mathrm{SD}}_{\mathrm{PROB}}$)		
Parameter	True Sd	Average	Standard deviation	% bias[a]	Relative efficiency[b]	Average	Standard deviation	% bias[c]	Relative efficiency[d]	Average	Standard deviation	% bias[e]
Total trials $\Sigma_i n_i = N = 15$, trials per level $n_i = 5$												
Model function 1: levels $x_i = -1, 0, 1$												
$\hat{m}$	0.481	0.474	0.162	−1.4	6.8	0.606	4.01	25.9	0.011	0.517	0.423	6.3
$\hat{g}$	0.395	0.376	0.041	−4.8	5.9	0.501	0.065	26.6	2.4	0.522	0.100	32.0
Total trials $\Sigma_i n_i = N = 30$, trials per level $n_i = 10$												
Model function 2: levels $x_i = -1, 0, 1$												
$\hat{m}$	0.296	0.322	0.159	8.6	1.4	0.331	0.973	11.8	0.038	0.291	0.191	−1.8
$\hat{g}$	0.371	0.352	0.043	−5.0	2.7	0.378	0.083	1.9	0.72	0.370	0.070	−0.2
Total trials $\Sigma_i n_i = N = 300$, trials per level $n_i = 100$												
Model function 3: levels $x_i = -1, 0, 1$												
$\hat{m}$	0.0824	0.0799	0.0090	−3.0	0.60	0.0819	0.0069	−0.61	1.04	0.0817	0.0070	−0.89
$\hat{g}$	0.107	0.105	0.0094	−1.7	0.24	0.108	0.0046	1.3	1.01	0.108	0.0046	0.69

Note. $m = 0$, $g = 1$, number of levels (l) = 3.
[a] % bias = $[(\mathrm{Ave}(\widehat{\mathrm{SD}}_{\mathrm{BOOT}}) - \mathrm{Sd})/\mathrm{Sd}] \cdot 100$, where Sd = "true Sd." [b] Relative efficiency = $\mathrm{Var}(\widehat{\mathrm{SD}}_{\mathrm{PROB}})/\mathrm{Var}(\widehat{\mathrm{SD}}_{\mathrm{BOOT}})$. [c] % bias = $[(\mathrm{Ave}(\widehat{\mathrm{SD}}_{\mathrm{INC}}) - \mathrm{Sd})/\mathrm{Sd}] \cdot 100$. [d] Relative efficiency = $\mathrm{Var}(\widehat{\mathrm{SD}}_{\mathrm{PROB}})/\mathrm{Var}(\widehat{\mathrm{SD}}_{\mathrm{INC}})$. [e] % bias = $[(\mathrm{Ave}(\widehat{\mathrm{SD}}_{\mathrm{PROB}}) - \mathrm{Sd})/\mathrm{Sd}] \cdot 100$.

by their smoothed values $\hat{y}_i$ estimated from the fitted curve of Equations 1a and 1b. Let $\hat{F}$ be the distribution with the rescaled binomial $\mathrm{Bi}(n_i, \hat{y}_i)/n_i$ at each level x_i, $i = 1, 2, \ldots, l$. Draw a bootstrap sample $(Y_1^*, Y_2^*, \ldots, Y_l^*)$ from $\hat{F}$ (the same size as the original data set) and fit the function of Equations 1a and 1b by maximizing the likelihood to obtain new estimates $\hat{m}^*$ and $\hat{g}^*$ (illustrated in Figure 1b and again in Figure 1c). Repeat this last step a large number B of times, to obtain B estimates $\hat{m}_1^*, \hat{m}_2^*, \ldots, \hat{m}_B^*$, and B estimates $\hat{g}_1^*, \hat{g}_2^*, \ldots, \hat{g}_B^*$. The bootstrap estimates of the standard deviations $\widehat{\mathrm{SD}}_{\mathrm{BOOT}}$ are given by the sample standard deviations

$$\widehat{\mathrm{SD}}_{\mathrm{BOOT}}(\hat{m}) = [\textstyle\sum_{b=1}^{B}(\hat{m}_b^* - \hat{m}_\cdot^*)^2/(B-1)]^{1/2},$$

$$\widehat{\mathrm{SD}}_{\mathrm{BOOT}}(\hat{g}) = [\textstyle\sum_{b=1}^{B}(\hat{g}_b^* - \hat{g}_\cdot^*)^2/(B-1)]^{1/2},$$

where $\hat{m}_\cdot^* = \sum_{b=1}^{B}\hat{m}_b^*/B$ and $\hat{g}_\cdot^* = \sum_{b=1}^{B}\hat{g}_b^*/B$. Figures 1d and 1e show histograms for 100 bootstrap replications from the sample set in Figure 1a, and the calculated bootstrap standard deviations. See Efron (1982) for further details.

Data Sets

The three methods for estimating standard deviations were each applied in 12 experimental conditions, with different numbers and spacings of the test levels x_i and numbers n_i of trials per level $i = 1, 2, \ldots, l$. For each experimental condition, 1,000 sample sets of data $(Y_1, Y_2, \ldots, Y_l)$ were generated. (Note that there were two levels of Monte Carlo: The 1,000 sample sets $(Y_1, Y_2, \ldots, Y_l)$ for each condition and the B bootstrap samples $(Y_1^*, Y_2^*, \ldots, Y_l^*)$ generated with each $(Y_1, Y_2, \ldots, Y_l)$ held fixed; Efron & Tibshirani, 1986.) With small data sets, there was an increased risk that estimated values of $\hat{m}$ and $\hat{g}$ would take extreme values; in particular, $\hat{m}$ could become infinite and $\hat{g}$ negative, zero, or positive infinite. Because extreme values would have had a destabilizing effect on the computation of the standard deviation, sample data sets yielding values of $\hat{m}$ or $\hat{g}^{-1}$ greater than 20 times the stimulus range were excluded, as a priori were those data sets that were degenerate, for example, when the Y_i were all identical or when the sets were of the form $(a, a, \ldots, a, b, b, \ldots, b)$, $0 \le a, b \le 1$. In an exhaustive analysis (Foster & Bischof, 1987) of one such case, where all 1,878 distinct positive pairs $\hat{m}$ and $\hat{g}$ were generated from Equation 1 with $m = 0$, $g = 1$, $x_i = -2, -1, \ldots, 2$, and $n_i = n = 5$, the proportion of pairs that was found to be inadmissible was 4.2%.

For each experimental condition, "true" values of the standard deviations, $\mathrm{Sd}(\hat{m})$ and $\mathrm{Sd}(\hat{g})$, were calculated by generating either 5,000 or 10,000 admissible data sets.

It should be noted that the parametric bootstrap and incremental methods may be applied to data sets in which the levels x_i are unequally spaced, the numbers n_i of trials at each x_i are unequal, and $n_i = 1$ for one or more x_i.

Performance of Standard Deviation Estimators

The principal measure of performance for the probit, incremental, and bootstrap estimators in each condition was the *percentage bias*, defined as the difference between the average of the estimate $\widehat{\mathrm{SD}}$ (taken over 1,000 samples) and the true value Sd, expressed as a percentage of the true value. For example, for the bootstrap estimate $\widehat{\mathrm{SD}}_{\mathrm{BOOT}}(\hat{m})$ of the standard deviation of the estimated midpoint $\hat{m}$, the percentage bias was

$$[(\mathrm{Ave}(\widehat{\mathrm{SD}}_{\mathrm{BOOT}}(\hat{m})) - \mathrm{Sd}(\hat{m}))/\mathrm{Sd}(\hat{m})] \cdot 100.$$

Table 2
Comparison of Bootstrap, Incremental, and Probit Estimators for the Standard Deviation of the Estimated Midpoint ($\hat{m}$) and Gradient ($\hat{g}$) for Model Function (Equation 1)

Parameter	True Sd	Bootstrap estimate $\widehat{SD}_{BOOT}$				Incremental estimate $\widehat{SD}_{INC}$				Probit estimate $\widehat{SD}_{PROB}$		
		Average	Standard deviation	% bias[a]	Relative efficiency[b]	Average	Standard deviation	% bias[c]	Relative efficiency[d]	Average	Standard deviation	% bias[e]
Total trials $\Sigma_i n_i = N = 25$, trials per level $n_i = 5$												
Model function 4:												
levels $x_i = -2, -1, 0, 1, 2$												
$\hat{m}$	0.356	0.330	0.089	−7.4	0.62	0.324	0.084	−9.1	0.70	0.337	0.070	−5.3
$\hat{g}$	0.319	0.298	0.033	−6.4	16.4	0.385	0.087	20.7	2.33	0.387	0.132	21.6
Model function 5:												
levels $x_i = -1, -0.5, 0, 0.5, 1$												
$\hat{m}$	0.357	0.389	0.265	9.0	5.5	0.497	2.17	39.2	0.082	0.366	0.621	2.6
$\hat{g}$	0.536	0.524	0.110	−2.3	1.5	0.554	0.198	3.3	0.46	0.472	0.135	−12.0
Total trials $\Sigma_i n_i = N = 50$, trials per level $n_i = 10$												
Model function 6:												
levels $x_i = -2, -1, 0, 1, 2$												
$\hat{m}$	0.246	0.229	0.037	−7.2	0.79	0.231	0.039	−6.2	0.74	0.232	0.033	−5.7
$\hat{g}$	0.285	0.281	0.048	−1.5	3.5	0.297	0.094	4.1	0.94	0.271	0.091	−4.7
Model function 7:												
levels $x_i = -1, -0.5, 0, 0.5, 1$												
$\hat{m}$	0.213	0.234	0.143	9.9	0.39	0.224	0.233	5.3	0.15	0.211	0.090	−1.1
$\hat{g}$	0.328	0.340	0.077	3.9	0.32	0.323	0.068	−1.4	0.41	0.309	0.043	−5.7
Total trials $\Sigma_i n_i = N = 500$, trials per level $n_i = 100$												
Model function 8:												
levels $x_i = -2, -1, 0, 1, 2$												
$\hat{m}$	0.0759	0.0723	0.0061	−4.7	0.26	0.0749	0.0031	−1.7	0.98	0.0748	0.0031	−1.4
$\hat{g}$	0.0744	0.0728	0.0087	−2.3	0.51	0.0738	0.0068	−0.9	0.82	0.0733	0.0062	−1.5
Model function 9:												
levels $x_i = -1, -0.5, 0, 0.5, 1$												
$\hat{m}$	0.0623	0.0599	0.0063	−3.9	0.58	0.0614	0.0048	−1.3	1.02	0.0615	0.0048	−1.3
$\hat{g}$	0.0938	0.0904	0.0076	−3.6	0.13	0.0932	0.0028	−0.6	0.96	0.0929	0.0027	−0.9

Note. $m = 0$, $g = 1$, number of levels (l) = 5.

[a] % bias = $[(\mathrm{Ave}(\widehat{SD}_{BOOT}) - \mathrm{Sd})/\mathrm{Sd}] \cdot 100$, where Sd = "true Sd." [b] Relative efficiency = $\mathrm{Var}(\widehat{SD}_{PROB})/\mathrm{Var}(\widehat{SD}_{BOOT})$. [c] % bias = $[(\mathrm{Ave}(\widehat{SD}_{INC}) - \mathrm{Sd})/\mathrm{Sd}] \cdot 100$. [d] Relative efficiency = $\mathrm{Var}(\widehat{SD}_{PROB})/\mathrm{Var}(\widehat{SD}_{INC})$. [e] % bias = $[(\mathrm{Ave}(\widehat{SD}_{PROB}) - \mathrm{Sd})/\mathrm{Sd}] \cdot 100$.

A second measure of performance was the *relative efficiency* of the bootstrap and incremental estimator with respect to the probit estimator, defined as the inverse ratio of the variances of the estimates. Hence, for the bootstrap estimate $\widehat{SD}_{BOOT}(\hat{m})$, the relative efficiency was

$$\mathrm{Var}(\widehat{SD}_{PROB}(\hat{m}))/\mathrm{Var}(\widehat{SD}_{BOOT}(\hat{m})).$$

Both $\widehat{SD}_{BOOT}(\hat{m})$ and $\widehat{SD}_{INC}(\hat{m})$ behaved as consistent estimators.

Procedure

For the probit method, maximum likelihood estimates were calculated by iterative regression, as described in Finney (1952, 1971), with a maximum of 50 cycles of the iteration and a convergence tolerance of 10^{-4}. For the incremental method, the partial derivatives in Equations 3a and 3b were each estimated by finite-difference approximations. The bootstrap estimates of the standard deviation were each based on 100 bootstrap replications ($B = 100$). (The effect of B on the variance of the bootstrap estimate of the standard deviation is considered later.)

For the incremental and bootstrap methods a nonlinear optimization technique modified from the simplex method (Nelder & Mead, 1965) was used to fit the model function to the data. Because of the sensitivity of the bootstrap standard deviation to occasional extreme values of $\hat{m}^*$ and $\hat{g}^*$, each distribution of $\hat{m}^*$ and $\hat{g}^*$ generated from a sample set was symmetrically two-fold Winsorized (Foster & Bischof, 1987).

Computations were carried out in FORTRAN on two mainframe computers, a Cyber 176 and a CDC 7600, each with floating-point precision of 15 significant decimal digits. The NAG routine G05EYF was used to generate pseudorandom integers (Numerical Algorithms Group, 1984). The two machines were used to spread the computational load, and in a number of control measurements produced identical results.

Results

The results of the Monte Carlo studies are shown in Tables 1–3 with the data grouped according to the number of stimulus levels and trials per level in the model psychometric functions.

Table 3
Comparison of Bootstrap, Incremental, and Probit Estimators for the Standard Deviation of the Estimated Midpoint ($\hat{m}$) and Gradient ($\hat{g}$) for Model Function (Equation 1)

		Bootstrap estimate $\widehat{SD}_{BOOT}$				Incremental estimate $\widehat{SD}_{INC}$				Probit estimate $\widehat{SD}_{PROB}$		
Parameter	True Sd	Average	Standard deviation	% bias[a]	Relative efficiency[b]	Average	Standard deviation	% bias[c]	Relative efficiency[d]	Average	Standard deviation	% bias[e]
Total trials $\Sigma_i n_i = N = 45$, trials per level $n_i = 5$												
Model function 10: levels $x_i = -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2$												
$\hat{m}$	0.250	0.233	0.050	−6.5	0.70	0.230	0.043	−7.7	0.92	0.235	0.042	−5.6
$\hat{g}$	0.361	0.378	0.125	4.5	1.1	0.404	0.192	11.6	0.45	0.296	0.129	−18.1
Total trials $\Sigma_i n_i = N = 90$, trials per level $n_i = 10$												
Model function 11: levels $x_i = -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2$												
$\hat{m}$	0.174	0.163	0.023	−6.3	0.71	0.167	0.019	−4.0	0.99	0.168	0.019	−3.2
$\hat{g}$	0.203	0.215	0.065	6.0	0.39	0.209	0.057	3.1	0.52	0.186	0.041	−8.4
Total trials $\Sigma_i n_i = N = 900$, trials per level $n_i = 100$												
Model function 12: levels $x_i = -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2$												
$\hat{m}$	0.0540	0.0517	0.0040	−4.4	0.19	0.0536	0.0018	−0.8	0.97	0.0536	0.0018	−0.8
$\hat{g}$	0.0546	0.0532	0.0051	−2.6	0.31	0.0545	0.0030	−0.2	0.87	0.0544	0.0028	−0.3

Note. $m = 0$, $g = 1$, number of levels (l) = 9.
[a] % bias = [(Ave($\widehat{SD}_{BOOT}$) − Sd)/Sd] · 100, where Sd = "true Sd." [b] Relative efficiency = Var($\widehat{SD}_{PROB}$)/Var($\widehat{SD}_{BOOT}$). [c] % bias = [(Ave($\widehat{SD}_{INC}$) − Sd)/Sd] · 100. [d] Relative efficiency = Var($\widehat{SD}_{PROB}$)/Var($\widehat{SD}_{INC}$). [e] % bias = [(Ave($\widehat{SD}_{PROB}$) − Sd)/Sd] · 100.

For each condition, summary data are shown for the bootstrap, incremental, and probit estimators. Figure 2 shows the effect of bootstrap replication number B on the variance of the estimators $\widehat{SD}_{BOOT}(\hat{m})$ and $\widehat{SD}_{BOOT}(\hat{g})$. The broken lines are linear least squares regressions.

Discussion

The bootstrap estimator was clearly the best estimator in each one of the three tables. In Table 1 the maximum magnitude of the percentage bias for the bootstrap estimator $\widehat{SD}_{BOOT}$ was 8.6%, for the incremental estimator $\widehat{SD}_{INC}$ 26.6%, and for the probit estimator $\widehat{SD}_{PROB}$ 32.0%; in Table 2, the maximum percentage biases were 9.9%, 39.2%, and 21.6%, respectively; and in Table 3, 6.5%, 11.6%, and 18.1%, respectively. The superiority of the bootstrap is most evident when the total number of trials in the sample is less than about 50 (model psychometric functions 1, 4, 5, and 10 in Tables 1-3), although the distribution of trials over levels was also important. The relative efficiency of the bootstrap estimator was also high when the total number of trials was small and generally exceeded the relative efficiency of the incremental estimator. Thus, in Model Function 1 (Table 1), where the total number of trials was 15, the relative efficiency of the bootstrap estimator for the standard deviation of the estimated midpoint was 6.8, and in Model Function 5 (Table 2), where the total number of trials was 25, it was 5.5. When the total number of trials was about 90 or more, all three estimators performed about the same, although, as is made clear later, the efficiency of the bootstrap could have been improved further.

The largest percentage biases in the probit estimator occurred in the estimation of the standard deviation of the estimated slope of the psychometric function. The performance of the probit estimator for both the standard deviation of the slope and of the midpoint may worsen with smaller numbers of trials or with asymmetric psychometric functions, ranging, for example, over 50–100% rather than over 0–100% (Foster & Bischof, 1987; McKee et al., 1985). Thus, in a separate simulation, Model Function 4 (Table 2) ranging over 50–100% yielded a bias for the probit estimator of the standard deviation of the estimated midpoint 4.4 times higher than that for the bootstrap estimator.

Effect of Replication Number B

It has been suggested that $B = 100$ is usually sufficient for estimating standard deviations (Efron, 1982; Hinkley, 1988). In Figure 2, the dependence of the variance of $\widehat{SD}_{BOOT}$ on $1/B$ is approximately linear. Increasing B from 100 to 200 reduced the variance in $\widehat{SD}_{BOOT}$ ($\hat{m}$) by about 40%, but increasing B again from 200 to 400 only reduced the variance by a further 20%. When the number of trials per level n_i was reduced (in Figure 2,

$n_i = 100$), the effect of B was found to be less important, and bootstrap relative efficiencies were generally higher (Tables 1–3). The decrease in $\widehat{\mathrm{SD}}_{\mathrm{BOOT}}$ with increase in B suggests that $\widehat{\mathrm{SD}}_{\mathrm{BOOT}}$ was being destabilized by a few outlying bootstrap replications not trapped by the Winsorization, and a more robust procedure may be preferred for the calculation of $\widehat{\mathrm{SD}}_{\mathrm{BOOT}}$.

In practice, when only modest numbers of data sets have to be analyzed rather than the many thousands considered here, it should be possible to afford large values of B. The efficiency of the bootstrap simulation itself may also be improved by incorporating variance-reduction techniques, including balanced sampling, which may lead to substantial reductions in the value of B for a given level of simulation error (Davison, Hinkley, & Schechtman, 1986; Hinkley, 1988). Hall (1989) has provided an analysis of three efficient bootstrap algorithms.

If confidence intervals rather than standard deviations were of interest, the minimum value of B would have to be increased by about a factor of 10 (Efron & Tibshirani, 1986). Some relevant methodological issues have been discussed by DiCiccio and Romano (1988), DiCiccio and Tibshirani (1987), Hall (1986), and Tibshirani (1988). Confidence intervals may be preferred when the bootstrap distribution is skewed or strongly non-normal and the standard deviation no longer provides a good indication of the precision of the point estimate.

Experimental Implications

When are standard deviation estimates of the kind considered here likely to be important? First, in some experiments it may be desirable to use no more than the minimum number of trials necessary to achieve a prescribed level of precision in a threshold estimate. Reliable standard deviation estimates are a prerequisite for such judgments and in medical assay are an essential adjunct to the specification of drug potencies in terms of mean-effective-dose (ED50) values. Second, in some psychophysical experiments, it may be difficult to repeat measurements. Thresholds may be changing rapidly, as in some sensory adaptation and recovery paradigms, or the total time available for obtaining data may be severely constrained, as in some clinical situations. Third, in such situations, estimates of the slope of a function and its precision may have diagnostic relevance for individual subjects. Although there have been suggestions to the contrary (e.g., Watson & Pelli, 1983), the slope of a psychometric function is not always invariant under changes in adaptation level, and a significant reduction in the magnitude of the slope may indicate pathology of sensory function (Patterson, Foster, & Heron, 1980). Slope precision is of course critical in medical assays when potency of a drug is being assessed in terms of the gradient of a dose–response relation (slope ratio assay; Finney, 1978). Fourth, even when repetition of measurements is feasible, estimates of the standard deviations of individual parameter estimates may still be useful in forming the best (minimum variance) estimate of the parameter, or in assessing the contribution of potential outliers to the mean. Finally, the magnitude of the estimated standard deviation may itself be used to decide among a number of competing parameters, such as midpoint, slope, and spread, each offering a summary of overall stimulus–response performance.

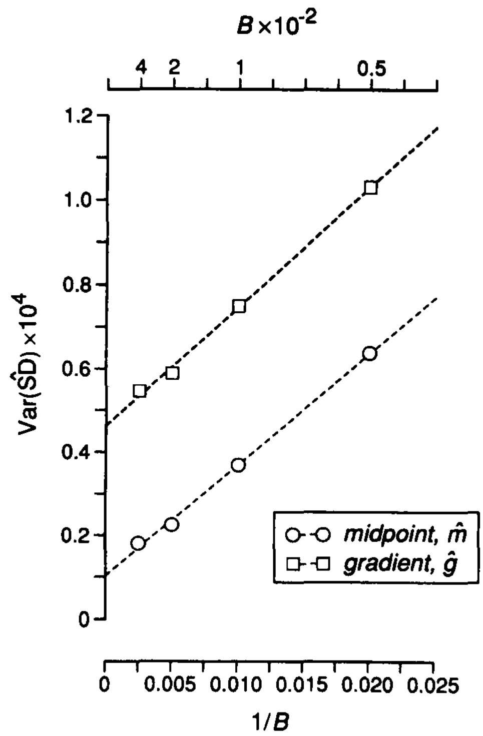

Figure 2. Variance of the bootstrap standard deviation estimate as a function of the bootstrap replication number B. (The data were generated from Equation 1 with $m = 0$, $g = 1$, $l = 5$, $n_i = n = 100$.)

The present analysis assumed a standard form for the psychometric function, a requirement imposed by the use of the traditional probit method. Suppose that the form of the psychometric function is unknown. Both the bootstrap and incremental methods can be used to obtain distribution-free estimates of the standard deviation of a threshold estimate, but, as Efron and Gong (1983) noted, a good parametric analysis, when appropriate, can be more efficient than the nonparametric counterpart. The smoothed versions of the bootstrap and incremental methods were introduced here to improve efficiency, but smoothing was not essential, and the variance of the estimates could have been reduced by some of the stabilization techniques cited earlier.

For large samples, the probit method is likely to continue as the method of choice, but, for medium-to-small samples, the use of formulae from classical asymptotic theory should be viewed with caution. In discussing maximum likelihood methods, Finney (1952, p. 246) was careful to emphasize that "the known optimal properties of maximum likelihood estimation relate to large samples, and some alternative may be superior in samples of finite size." The simulations that were presented here were intended to span a representative variety of data sets of finite size that might occur in adaptive or fixed-levels designs (method of constant stimuli). Because of the effects of stimulus-level spacing it is not possible to give a general lower limit on sample size for which probit analysis gives inappropriate standard deviation estimates. A conservative recommendation might be to consider use of the bootstrap method as an alternative when the total number of trials falls somewhat below 100, but this figure may have to be revised upward when the psychometric function is asymmetric or the spacing of test levels is not optimum.[3]

[3] A FORTRAN listing of the main programs used in this study is available on request from David H. Foster.

References

Cox, D. R. (1970). *The analysis of binary data.* London: Methuen.

Davison, A. C., Hinkley, D. V., & Schechtman, E. (1986). Efficient bootstrap simulation. *Biometrika, 73,* 555–566.

DiCiccio, T. J., & Romano, J. P. (1988). A review of bootstrap confidence intervals. *Journal of the Royal Statistical Society* (Series B), *50,* 338–354.

DiCiccio, T., & Tibshirani, R. (1987). Bootstrap confidence intervals and bootstrap approximations. *Journal of the American Statistical Association, 82,* 163–170.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans.* (CBMS-NSF Regional Conference Series in Applied Mathematics, No. 38.) Philadelphia, PA: Society for Industrial and Applied Mathematics.

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

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science, 1,* 54–75.

Emerson, P. L. (1984). Observations on a maximum likelihood method of sequential threshold estimation and a simplified approximation. *Perception & Psychophysics, 36,* 199–203.

Falmagne, J.-C. (1982). Psychometric functions theory. *Journal of Mathematical Psychology, 25,* 1–50.

Fechner, G. T. (1860). *Elemente der Psychophysik.* Leipzig: Breitkopf und Härtel.

Finney, D. J. (1947). *Probit analysis* (1st ed.). Cambridge, England: Cambridge University Press.

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

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

Finney, D. J. (1978). *Statistical method in biological assay.* London: Charles Griffin.

Foster, D. H. (1986). Estimating the variance of a critical stimulus level from sensory performance data. *Biological Cybernetics, 53,* 189–194.

Foster, D. H., & Bischof, W. F. (1987). Bootstrap variance estimators for the parameters of small-sample sensory-performance functions. *Biological Cybernetics, 57,* 341–347.

Hall, P. (1986). On the number of bootstrap simulations required to construct a confidence interval. *The Annals of Statistics, 14,* 1453–1462.

Hall, P. (1989) On efficient bootstrap simulation. *Biometrika, 76,* 613–617.

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

Lindley, D. V. (1965). *Introduction to probability and statistics from a Bayesian viewpoint, part 1: Probability.* (pp. 134–136). Cambridge, England: Cambridge University Press.

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.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal, 7,* 308–313.

Numerical Algorithms Group. (1984). *FORTRAN library manual, Mark 11* (Vol. 5). Oxford, England: Author.

Patterson, V. H., Foster, D. H., & Heron, J. R. (1980). Variability of visual threshold in multiple sclerosis: Effect of background luminance on frequency of seeing. *Brain, 103,* 139–147.

Shelton, B. R., Picardi, M. C., & Green, D. M. (1982). Comparison of three adaptive psychophysical procedures. *Journal of the Acoustical Society of America, 71,* 1527–1533.

Taylor, M. M., Forbes, S. M., & Creelman, C. D. (1983). PEST reduces bias in forced choice psychophysics. *Journal of the Acoustical Society of America, 74,* 1367–1374.

Tibshirani, R. (1988). Variance stabilization and the bootstrap. *Biometrika, 75,* 433–444.

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

Received September 19, 1989
Revision received March 15, 1990
Accepted April 6, 1990 ■