Perception & Psychophysics
1985, 37 (4), 286-298

Statistical properties of forced-choice psychometric functions: Implications of probit analysis

SUZANNE P. McKEE
Smith-Kettlewell Institute of Visual Sciences, San Francisco, California

STANLEY A. KLEIN
College of Optometry, University of Houston, Houston, Texas

and

DAVIDA Y. TELLER
University of Washington, Seattle, Washington

Perception & Psychophysics
1985, 37 (4), 286-298

Statistical properties of forced-choice psychometric functions: Implications of probit analysis

SUZANNE P. McKEE
Smith-Kettlewell Institute of Visual Sciences, San Francisco, California

STANLEY A. KLEIN
College of Optometry, University of Houston, Houston, Texas

and

DAVIDA Y. TELLER
University of Washington, Seattle, Washington

Probit analysis was applied to the problem of threshold estimation from psychometric functions derived from the two-alternative forced-choice (2AFC) method of constant stimuli. Threshold estimates from 2AFC experiments are surprisingly poor: They are about twice as variable as corresponding estimates based on the traditional yes-no method of constant stimuli, and their asymmetrical confidence limits are not readily predicted from conventional standard error formulas. All of these faults are exacerbated in small samples. Computer simulations demonstrated that, for small samples, the probit analysis equations do not give a valid estimate of threshold variability. The variability of staircase estimates of threshold cannot be less than the variability of threshold estimates derived from the method of constant stimuli given an optimum placement of trials. Hence our findings also define the minimum variability of all staircase estimators under the assumptions of probit analysis.

Forced-choice techniques, in combination with the method of constant stimuli, are increasingly common in modern psychophysical studies. In a typical two-alternative forced-choice (2AFC) experiment, a stimulus is presented in one of two possible positions on each of a series of trials, and the subject judges the position of the stimulus on each trial. Several different stimulus levels, varying along some physical dimension (such as intensity), are presented in random order for a substantial number of trials each. In an orderly data set, the subject's percent correct will vary from near 50% (chance) for stimuli too weak to be detected to near 100% for stimuli that are readily detected. Some empirical or theoretical curve is fitted to the data and used to estimate one or more parameters of the assumed underlying population. The most commonly estimated value is the *threshold*, T_{75}, which is the stimulus value needed for the subject to be correct 75% of the time. This threshold depends on the *location* along the abscissa of the whole psychometric function. The *slope* of the function may also be of interest. If the cumulative normal curve is used as the theoretical function, the threshold, T_{75}, corresponds to the mean, μ, and the slope, β, corresponds to the reciprocal of the standard deviation, σ, of the normal curve (i.e., $\sigma = 1/\beta$).

With normal adult subjects, it is feasible to run 100 or more trials for each stimulus level. Then error variance is small, the location and slope of the psychometric function can be judged by eye, and typically no elaborate curve-fitting or statistical analyses are needed. However, with less docile subjects, such as infants and clinical patients, the number of trials may be restricted, and the statistical properties of threshold estimates derived from small samples become important. These properties are surprisingly weak: Threshold estimates derived from forced-choice data may have unacceptably large standard errors, and the confidence limits may be asymmetrical and significantly larger than would be naively predicted

This work was supported by NIH Grants 5 P-30-EY00186, R01-EY03976, R01-EY04776, and R01-EY02920 and by the Smith-Kettlewell Eye research Foundation. We thank Polly Feigel, Lou Godio, Jacob Nachmias, Walter Makous, and Roger Watt for their helpful critical commentary on an earlier version of this manuscript, and Donald MacLeod for useful conversations on this topic. Gerald Westheimer wrote the original probit program for the yes-no case. This program served as a model for subsequent programs that employed different strategies to handle the 2AFC case and the simulations. We also want to acknowledge considerable assistance from Martha Teghtsoonian and an anonymous reviewer in the editing of the final version. A computer program for 2AFC probit analysis is available from the first author on written request. Another program, available from the second author, works with several shapes of the psychometric function; it has several options for estimating confidence limits and an option for estimating the upper asymptote.

S. P. McKee's mailing address is: Smith-Kettlewell Institute of Visual Sciences, 2232 Webster St., San Francisco, CA 94115.

from the standard error. More than 100 trials are required before confidence limits become well behaved.

The purpose of this paper is fourfold: (1) to explicate the reasons for these statistical properties at a simple graphical level; (2) to describe a common statistical approach—probit analysis (Finney, 1971)—that is often applied to yes-no psychophysical data, and to explain its use in the 2AFC case; (3) to check the validity of the probit analysis equations by computer simulation of the sampling distribution of the statistic T_{75}; (4) to explore the implications of these results for the use of 2AFC techniques.

GRAPHICAL ANALYSIS

This section provides the reader with an intuitive understanding of the statistical properties of estimates of T_{75} derived from 2AFC data, given the assumptions of probit analysis. Although the results of this graphical analysis are inexact in certain respects (see the appendix), this approach will introduce several concepts that are important to later sections of this paper. We will begin with an explanation of the notation used on the axes of these 2AFC graphs.

On each trial in the traditional method of constant stimuli as applied to a detection task, the subject is shown one sample from a set of stimuli that vary along some physical dimension and is asked whether the presented sample exceeds his criterion ("yes or no?"). Data from many trials are used to determine the stimulus level needed to produce a particular positive response percentage, usually 50% "yes," corresponding to the midpoint of the psychometric function. Because the psychometric function generated by this yes-no procedure often resembles a normal ogive, probit analysis can be used to estimate the statistical properties of these thresholds.

Probit analysis can also be used to estimate threshold statistics from 2AFC data by rescaling the cumulative normal function to extend from a lower asymptote C to an upper asymptote D. In an "ideal" 2AFC task, the percentage correct varies from 50% to 100% so that C = 0.5 and D = 1.0. C and D are incorporated into probit analysis by assuming an underlying function in which the probability (P) goes from 0 to 1.0, and, using Abbott's formula to obtain a probability (or percentage correct), P*, whose limits are C and D:

$$P^* = C + (D-C)\,P. \qquad (1)$$

For the ideal 2AFC case, where C = 0.5 and D = 1.0, the midpoint of the psychometric function occurs at P* = 0.75 and the corresponding stimulus value is designated T_{75}.

The top row of Figure 1 shows the two idealized psychometric functions for yes-no and 2AFC. The right-hand ordinate of Figure 1B gives the probabilities (P) of the underlying cumulative normal function that have been rescaled to fit the left-hand ordinate P*, the percentage correct. Usually, the abscissa of a graphed psychometric function is scaled in units of the stimulus parameter. Here, the abscissa is scaled in units of the standard deviation, σ, of the cumulative normal function, commonly called Z-score units. Thus, the stimulus values of −1, 0, and 1 correspond to the percentages correct (P*) of 58%, 75%, and 92% for the 2AFC function and percentages "yes" (P) of 16%, 50%, and 84% for the yes-no function.

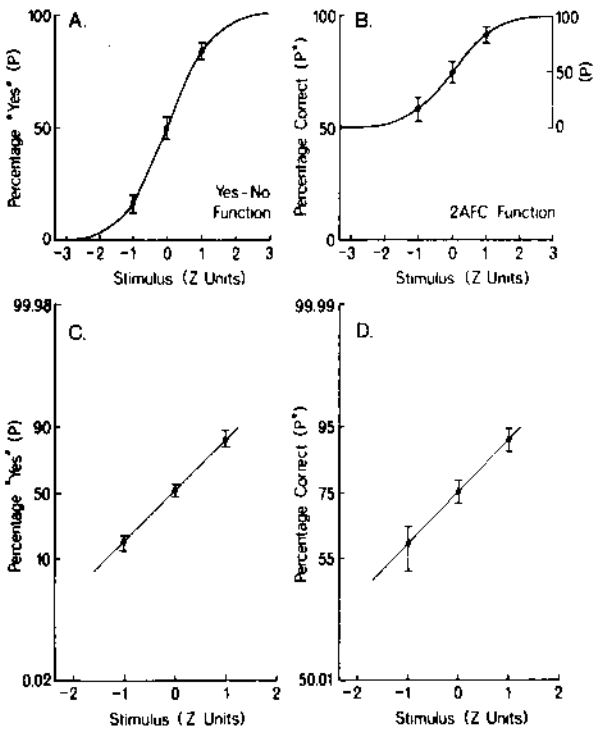

Figure 1. Graphical illustration of cumulative normal curves and binomial variability, for both yes-no (A and C) and 2AFC (B and D) cases, on both linear (A and B) and probability (C and D) ordinates. In each case, error bars represent ±1 standard error for n = 100. The abscissa is scaled in Z-units, that is, units of σ, the standard deviation of the cumulative normal curve. In C and D, equal distances on the ordinate represent equal standard deviations of the cumulative normal curve, so that the cumulative normal is converted into a straight line. Zero on the abscissa corresponds to the classical threshold, that is, P = 0.5 for the yes-no case and P* = 0.75 for the 2AFC case. The hypothetical stimuli are placed at −1, 0, and 1 Z-unit; this condition represents the "standard case" for the remainder of this paper.

Standard errors for n = 100 trials per sampling point have been plotted on both functions at the stimulus values of −1, 0, and +1. These standard errors are calculated from the usual formula for a standard error of a proportion for random samples of size n drawn from a binomial distribution: $\sqrt{PQ/n}$ for the yes-no case or $\sqrt{P^*Q^*/n}$ for the 2AFC case, where Q = 1 − P and Q* = 1 − P*.

For yes-no data, the sampling variability is largest near 50% and is symmetrically smaller as one moves toward the lower and upper asymptotes. But forced-choice data span only the range from P* = .5 upward. The sampling variability is largest at C and smallest at D. Forced-choice

data thus have a strong asymmetry, with the data points near 50% being likely to deviate more from their corresponding population values than do points near 100%.

When a theoretical curve is fitted to the data by probit analysis, the individual data points are weighted inversely with their intrinsic binomial variability. Thus, in the 2AFC case, data points sampled from the lower part of the function will be weighted less heavily, and hence constrain the fitted curve less, than data points sampled from the upper part of the function.

A second factor influences the degree of constraint exerted by a given data point on the curve-fitting process—the slope of the normal ogive at that point. Consider translating a normal ogive laterally until the curve intersects the upper or lower end of the error bars. Given error bars of a fixed size, the curve could be moved over a much larger range where the slope is shallow than where it is steep. Points far in the tail may have much smaller binomial variability than those near the midpoint, yet their influence on the curve-fitting process may be negligible. For forced-choice data, points near 50% constrain the fitting of the curve very litle indeed, since they suffer both from large error bars and minimal slope.

Linearizing Transformations

The fitting of a curvilinear function to a set of data is simplified if the function can be converted into a straight line. In psychophysical practice, data are often plotted on "probability paper," transforming a cumulative normal curve into a straight line. Equal distances on the ordinate are equal standard deviations of a normal distribution (equal Z-units). Linearized versions of Figures 1A and 1B are shown in Figures 1C and 1D. For these curves, the abscissa is also scaled in Z-units, so the slope of the lines is 1.0.

When the cumulative normal ogive is transformed into a linear function, differences in the value of the slope are expressed by differential "stretching" along the ordinate. The ends of the error bars are tied to the ordinate and stretch along with it; thus the relative lengths of the error bars change. For the transformed yes-no case, the error bars are *smallest* near 50% and *largest* in the tails. In the 2AFC case, the smallest error bars are at 83% and the largest errors are found in the lower tail, where the influence of a shallow slope, now translated into a magnified distance on the ordinate, is superimposed on the already large binomial variation. The intrinsic variability of these functions depends on the magnitude of these error bars. Given the same number of trials, the variability of threshold estimates derived from 2AFC data is greater than the variability of thresholds based on "yes-no" data provided that both types of data are adequately described by cumulative normal functions with the same value of σ.

Confidence Limits

The 2AFC case is examined further in Figure 2, where the vertical error bars now represent the 95% confidence limits, ± 1.96, $\sqrt{P^*Q^*/n}$. Throughout this paper, n is the number of trials per point, k is the number of stimuli, and N = nk is the total number of trials. Figure 2 shows the effect of varying n (100 vs. 40) and consequently N (300 and 120) for three stimulus values. The graphical approach for two cases—*fixed* (or *known*) *slope* and *variable* (or *unknown*) *slope* are shown in the left and right columns of Figure 2, respectively. In the fixed-slope case, the assumed slope equals 1 and only the location parameter is to be estimated; in the variable-slope case, both the location and the slope are to be estimated.

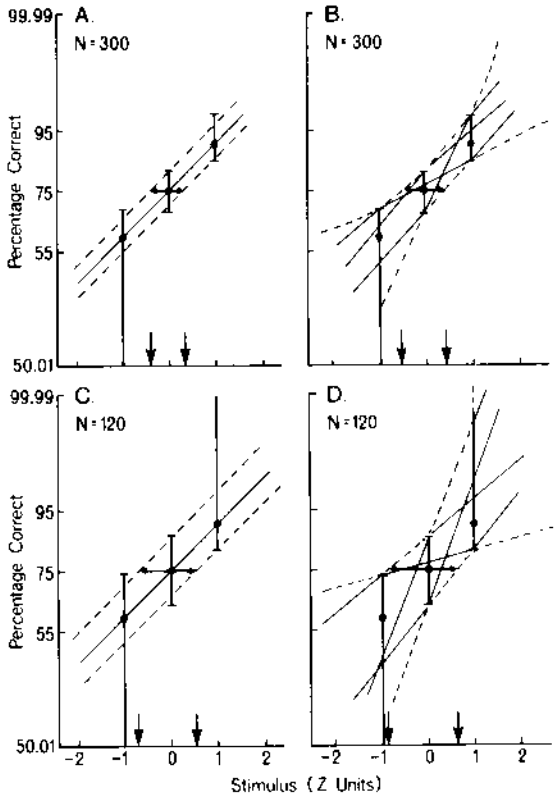

Figure 2. Graphical approximation of confidence limits for the estimate of T_{75} in the 2AFC case, as influenced by the number of trials and the assumption of a fixed (or known) slope vs. a variable (or unknown) slope. A—N=300, fixed slope; B—N=300, variable slope; C—N=120, fixed slope; D—N=120, variable slope. Error bars represent ±1.96 binomial standard errors for each point. The dotted lines represent the outer limits of straight lines that can be fitted through the error bars. The horizontal arrow through the data point at Z=0 spans the distance between the dotted lines, and represents a graphical estimate of the confidence limits. The two arrows above the abscissa mark the estimated confidence limits.

In the fixed-slope cases, the outer dotted lines delimit the family of possible lines of slope = 1 that will fall entirely within the binomial error bars for the three chosen stimulus values. Similarly, in the corresponding variable-slope cases, the outer dotted lines delimit the family of all possible straight lines, of whatever slope, that will fall entirely within the same three error bars. The arrows dropped from the limiting dotted lines to the abscissa provide graphical estimates of the 95% confidence limits for T_{75}. It must be emphasized that these graphical estimates

of the confidence limits are inexact (too large) and are presented only as an intuitive guide.[1]

Several inferences can be drawn from Figure 2. First, in each pair of graphs, confidence limits for T_{75} are larger in the variable-slope case. The need to estimate the slope parameter results in greater uncertainty in the estimate of T_{75} than does estimating the location parameter alone. Second, obviously, the confidence limits become larger with smaller values of N. Third, the confidence limits are asymmetrical. This asymmetry also implies that the confidence limits cannot readily be calculated from commonly used (symmetrical) formulas such as ±1.96 standard errors, and that intuitive statistical comparisons based on values of the standard error will be misleading.

The optimum placement of trials along the stimulus continuum will vary somewhat for different values of N, and with the choice of a fixed or variable-slope approach. When the slope is known, it makes sense to place all trials near the point that provides maximal information on the location of the curve, and this point is the same (83%) for all values of N. When the slope is unknown, the choice of stimulus locations involves a careful balancing of two factors—the magnitudes of the error bars, which are themselves asymmetrical about T_{75}, and the vertical separation between the tested points—in order to minimize the confidence limits for the estimated T_{75}. The balance will vary with N: the smaller N, the farther displaced from the center of the distribution will be the optimal stimulus values.

Asymmetrical Placement of Trials

Since the binomial errors associated with a 2AFC psychometric function are asymmetrical, sampling above the center of the distribution is generally a better choice than sampling below its center. This effect is illustrated in Figure 3 for the variable-slope case.

Figures 3A and B show, for n = 100 trials and a stimulus spacing of 1, the effect of choosing the stimuli on the high side of the distribution. The asymmetrical choice diagramed in Figure 3A slightly increases the confidence interval for T_{75}, relative to the symmetrical case shown in the previous figure (2B). Note, however, that, if the upper asymptote were less than 1.0, the confidence limits could increase dramatically. There are some asymmetrical choices, such as the one shown in Figure 3B, that produce estimates of T_{75} with a variance nearly equal to symmetrical sampling. On the other hand, sampling much below the center of the function can have disastrous effects on the variance, as shown in Figure 3D.

In summary, graphical analysis illustrates the following properties of 2AFC psychometric functions:

(1) The variability of the estimated threshold (T_{75}) in the 2AFC case is greater than the variability of the estimated threshold (T_{50}) in the yes-no case.

(2) Estimating both the threshold and the slope introduces greater uncertainty in the estimate of T_{75} than does estimating the threshold when the slope is assumed to have some fixed value.

(3) Although the choice of symmetrically placed stimuli will often produce the smallest estimated variance of T_{75}, a range of stimuli placed somewhat asymmetrically above T_{75} will work as well (if the upper asymptote is near 100%), whereas displacement of the stimuli much below T_{75} leads to a deterioration of the estimates of T_{75}.

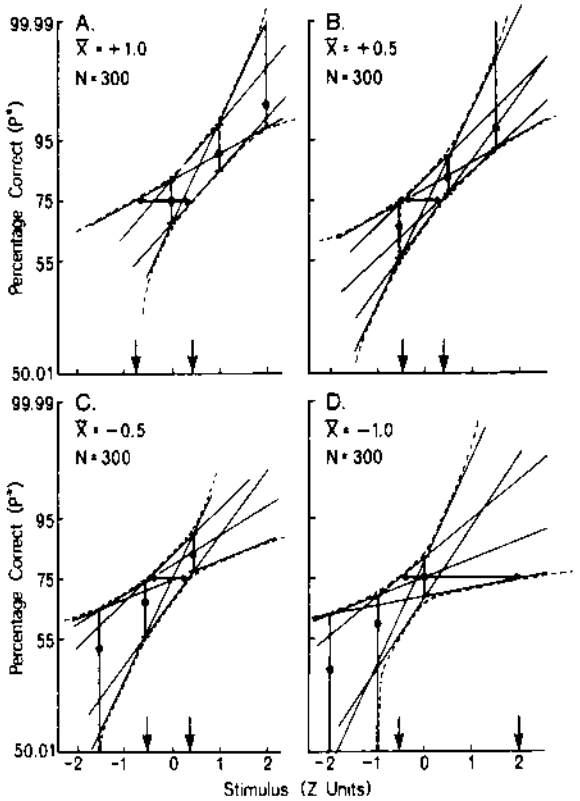

Figure 3. Graphical illustration of the effect of stimulus placement on the confidence limits of T_{75} estimated from 2AFC data. All lines and symbols as in Figure 2. Sampling above T_{75} (A and B) produces only a small increase in the uncertainty of the estimated T_{75}, and is a better strategy than sampling below T_{75} (C and D).

PROBIT ANALYSIS APPLIED TO THE 2AFC EXPERIMENT

Probit analysis is an iterative procedure for fitting a cumulative normal curve to a set of data, estimating the best choice for the parameters of the function according to a maximum likelihood criterion. If there are sufficient degrees of freedom in the data set and a sufficient number of trials, all parameters associated with the psychometric function (μ, σ, C, D) can be estimated (see Finney, 1971, chap. 7), but for purposes of the present discussion we assume that C and D are known, so that only estimates of μ and σ are required.

In the computational scheme described by Finney (1971), the observed probabilities are initially transformed into Z-units (cf. Figure 1) and a provisional line is fit to the data either by eye or through some more quantitative approach such as a least squares estimate of the y-intercept

and the slope. Since observations with smaller error bars (Figure 1) reduce uncertainty about the location and slope of the best-fitting function more than do observations with large error bars, differential weights are then assigned to the points on the provisional line. The weighting coefficients, w, depend directly on the slope of the cumulative normal function at the tested stimuli and inversely on P*Q*. Each point is also weighted in direct proportion to n, the number of trials for that stimulus. The probit calculation can now employ the statistical structure of weighted linear regression to estimate a best-fitting function. The regression procedure is performed repeatedly so that successive estimates of the parameters converge on the maximum likelihood estimates of the y-intercept, α, and the slope, β. These parameters are simply related to the mean and standard deviation of the normal function: $\mu = -\alpha/\beta$ and $\sigma = 1/\beta$.

Standard Errors

An attractive feature of probit analysis is that it can provide quantitative estimates of the standard error of estimation of T_{75}. In probit analysis, the simplest analytic formula for the standard error of the mean (Finney, 1971, p. 33) is

$$\frac{1}{b\sqrt{\Sigma nw}}, \qquad (2)$$

where b is the sample estimate of the slope β of the transformed function, and Σnw is the sum of the products of the probit weights w and the number of trials n for each tested stimulus.

This formula is a variation of the common statistical formula $s/\sqrt{N}$ used to estimate the standard error of the mean, because b is the reciprocal of the sample standard deviation, s, and Σnw, the weighted sum of the number of trials at each point, is used in place of N. In other words, in probit analysis, as in other realms of parametric statistics, the standard error of estimation of the mean depends directly on the sample standard deviation, s, and inversely on the square root of the number of trials. Use of this simple "fixed slope" formula for the standard error will result in a serious underestimation of the true value of the standard error if the total number of trials is small, or if the sampled observations are not centered near T_{75}. A better estimate of the standard error is given by the following "variable-slope" formula (Finney, 1971, p. 34):

$$SE = 1/b\sqrt{\frac{1}{\Sigma nw} + \frac{(T_{75}-\bar{x})^2}{\Sigma nw(x-\bar{x})^2}}, \qquad (3)$$

where $\bar{x}$ is the weighted mean of the sampled stimulus array, and $1/\Sigma nw(x-\bar{x})^2$ is the variance of the slope.

Predictably, the most important parameter controlling the magnitude of the standard error is sample size. Using Equation 3, we calculated the standard error as a function of the total number of trials (N) for one condition, which we call the standard case (see Figures 1A and 1B). For these calculations, we assumed a cumulative normal curve with $\mu = 0$ and $\sigma = 1$, so the slope of the function is 1. Thus, for the standard case, the sampled stimuli fall at -1, 0, and +1 Z-units, corresponding to percents correct (P*) of 58%, 75%, and 92% for 2AFC and to percents yes of 16%, 50%, and 84%, for yes-no. Because the stimulus values are given in Z-units, the calculated standard error will also be in Z-units. For example, if the standard error is 1 Z-unit, then the uncertainty about the location of T_{75} extends over most of the region covered by the psychometric function, that is, from the stimulus value corresponding to 58% to the value corresponding to 92%.

In Figure 4, we have plotted the calculated standard errors for the yes-no and 2AFC techniques as a function of N, along with a line that falls as $\sqrt{N}$. The line is displaced rightward by a factor of about 4 for the 2AFC case; about 4 times as many trials must be used in 2AFC as in yes-no to achieve the same value of the standard error. For any particular sample size, assuming that the value of σ remains constant across changes in psychophysical technique, the standard error for 2AFC is roughly twice the size of the standard error found with yes-no techniques. In a sense, this relationship is expected because the slope of the 2AFC technique is half the value of the slope for the yes-no technique, as is apparent when Figures 1A and 1B are compared.

Sampling Strategies

Clearly, the experimenter wishes to sample the stimulus domain in a way that provides unbiased, minimum variance estimates of T_{75}. In principle, the experimenter can choose the range, the number of sampled stimuli, and the region of the stimulus domain sampled with respect to the psychometric function. To compare sampling strategies, we compute the standard error of T_{75} for different ranges (R), different numbers of stimuli (k), and different regions of the stimulus domain, as specified by the center ($\bar{X}$) of the sampled stimulus set. These choices are diagramed in Figure 5. Calculations were performed with 60, 120,

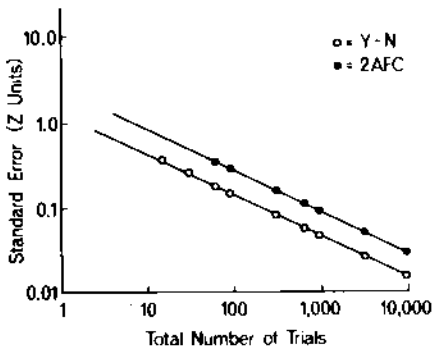

Figure 4. The standard error as a function of the total number of trials for the standard case of Figure 1. The continuous lines show predicted change if standard errors decreased as a function of $\sqrt{N}$. The standard error of T_{75} estimated from 2AFC data is about twice the size of the standard error of T_{50} estimated from yes-no technique, assuming the same value of σ for both techniques.

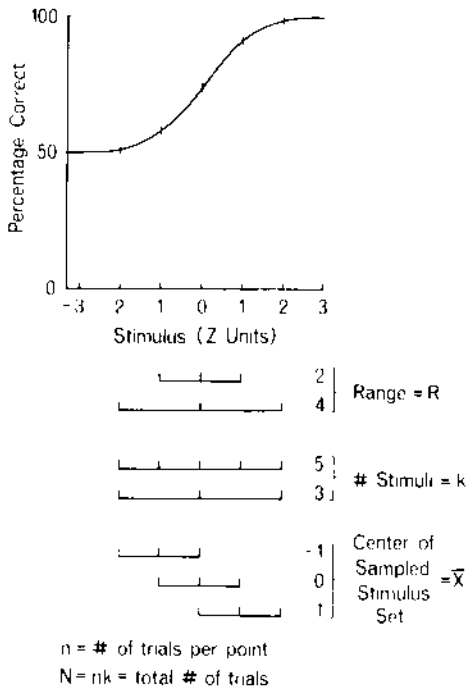

Figure 5. Graphical illustration of possible sampling strategies. In principle, experimenters can control the range (R), the number of stimuli (k), and the placement of the stimuli with respect to the psychometric function ($\bar{X}$).

and 300 total trials in the sample. Over a reasonable range of variations, manipulations of the value of R, k, and $\bar{X}$ had a relatively small effect on the value of the standard error. In the variable slope case, assuming a reasonable choice of stimulus placement, the standard error for 2AFC is approximately:

$$3\sigma/\sqrt{N}.$$

Confidence Limits

In most statistical contexts, when a sample size of 30 or more observations is used, the 95% confidence limits of the mean extend from about 2 standard errors below, to about 2 standard errors above, the mean. This simple assumption often guides inferences about the significance of the differences observed between experimental conditions. In the case of 2AFC experiments, such an estimate may be misleading, because, as the graphical analysis suggests, the true confidence limits are sometimes large and are frequently asymmetrical, with the lower limit usually being farther from T_{75} than the upper limit.

Finney recommends the use of the general formula given below to calculate the fiducial limits for T_{75}:

$$T_{75} + \frac{g}{1-g}(T_{75}-\bar{x}) \pm \frac{t}{b(1-g)}\sqrt{\frac{1-g}{\Sigma nw} + \frac{(T_{75}-\bar{x})^2}{\Sigma nw(x-\bar{x})^2}}, \quad (4)$$

where t is the normal deviate corresponding to the desired limits, for example, t = 1.96 for the 95% fiducial limits, and

$$g = \frac{t^2}{b^2\Sigma nw(x-\bar{x})^2}.$$

We have calculated the 95% fiducial limits using this equation for many combinations of stimulus placement and sample size (Teller, 1985). In many cases, the calculated 95% fiducial limits estimated by Equation 4 were very large. For N ranging between 60 and 100 trials, the fiducial limits extended from 2 to 4 Z-units in size, and for N smaller than 50-60 trials, the fiducial limits were indeterminate, even for optimal stimulus placement. Fiducial limits are indeterminate when the value of g exceeds 1.0, and g will generally exceed 1.0 for small samples ($N < 60$), using the 2AFC experimental procedure.

SIMULATIONS

Our major concern in this paper has been an adequate description of the variability of 2AFC thresholds, particularly for small samples. To check the validity of the standard error estimated from Equation 3 and the fiducial limits estimated from Equation 4, we used two computer-simulation techniques to examine the sampling distribution of T_{75}.

In the first technique, enumeration, all possible outcomes for each of the stimuli were weighted by the binomial probability that such an outcome could occur. For example, if the number of stimuli were 2, with 30 trials each, then the possible outcomes for each stimulus included: 0 correct, 30 wrong; 1 correct, 29 wrong; 2 correct, 28 wrong, etc. Each of the possible combinations of outcomes for the two stimuli defined a psychometric function and a threshold estimate, T_{75}. In the second technique, Monte Carlo simulation, the computer generated a sample percent correct for each value of P* by calling a random-number generator biased at the P* value n times. Probit analysis was then used to fit a best-fitting cumulative normal curve to the computer-generated data and to provide an estimate of T_{75}. This sequence was repeated 1,200 times to yield 1,200 estimates of T_{75}. For both simulation techniques, the upper and lower threshold values that would exclude the upper and lower 2.5% of the simulated population, respectively, were taken as the 95% confidence limits.

We first simulated the sampling distribution of T_{75} for the case of the two stimuli and a sample size of 60 trials (30 trials per stimulus), exploring the effect of range and stimulus placement on the confidence limits. Figure 6 shows the 95% fiducial limits (× symbols) calculated from Equation 4, and the 95% confidence limits based on the simulated distributions from the enumeration technique (squares) and from the Monte Carlo technique (filled circles). The fiducial limits from Equation 4 were indeterminate for all other values of stimulus placement ($\bar{X}$) for which no points are plotted.

All three methods were in rough agreement on the magnitude of the upper confidence (or fiducial limit),

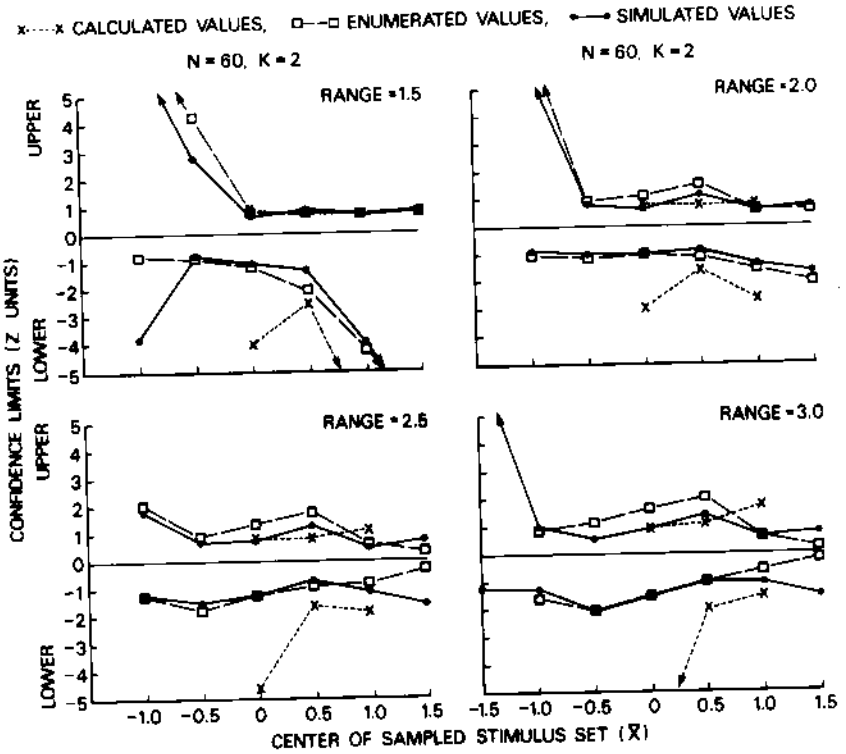

Figure 6. Confidence limits of the estimated value of T_{75} for N = 60, k = 2 for different ranges (R = 1.5, 2.0, 2.5, 3.0) as a function of stimulus placement as indicated by the center of sampled set ($\bar{X}$). The dotted line (×) shows the confidence limits calculated from equation (4). The dashed line (□) shows the simulated confidence limits found with the enumeration technique. The continuous line (•) shows the simulated confidence limits found with the Monte Carlo technique.

which hovered at a value of about twice the calculated standard error for values of $\bar{X}$ near the center of the psychometric function or above it. The lower fiducial limit from Equation 4 was typically much larger than the simulated values. We concluded that this equation failed to give an accurate estimate of the confidence limits of small 2AFC samples.

Equation 4 is based on an approximation that could be inappropriate for small samples and a high value of C, for example, 0.5 for the 2AFC case. Finney (1971) states that well-behaved data almost always give a value of g substantially smaller than 1.0 and usually less than 0.4. For the 2AFC case, we found that the value of g generally exceeds 0.4 when the total number of trials is less than 130, and often is greater than 1.0 even for optimally placed stimuli when N is less than 60 trials.

The patterns of results from the two simulation techniques were quite similar, although generally the Monte Carlo technique produced somewhat smaller values.[2] The confidence limits estimated by enumeration are a more accurate representation of the true characteristics of the underlying sampling distribution of T_{75}. However, the enumeration approach sometimes led to counterintuitive conclusions. Consider the confidence limits from the enumerations shown in the right-hand corner of Figure 6 for R = 3 and $\bar{X} = 1.5$. The true percents correct for the two stimuli falling at 0 and 3 corresponded to 75% and 99.9%, respectively, and, therefore, the most common value associated with the stimulus at 3 Z-units was 30 correct, 0 wrong. Mathematically, 100% correct is infinitely far from the center of a cumulative normal distribution. Thus, the psychometric function for this enumerated condition was often infinitely steep. The estimated value of T_{75} was the stimulus value corresponding to the other percent correct, which in this case was 0, since the best-fitting function frequently consisted of a vertical line through 0. The residual variation in T_{75} depended only on those cases in which the percent correct at 3 Z-units was less than 100%, so the confidence limits estimated by enumeration for this point were very small.

This peculiar case revealed a danger in small-sample-variable-slope estimation, particularly for k = 2. As the number of trials per stimulus is decreased, the chance of infinitely steep functions is increased. If, for example, only 10 trials had been used for each of two stimulus levels corresponding to true percents correct of 58% and 92%, then more than 80% of the functions would be infinitely steep, because one of the stimuli would have produced a percent equal to or below 50% or equal to 100%. Infinitely steep psychometric functions contain lit-

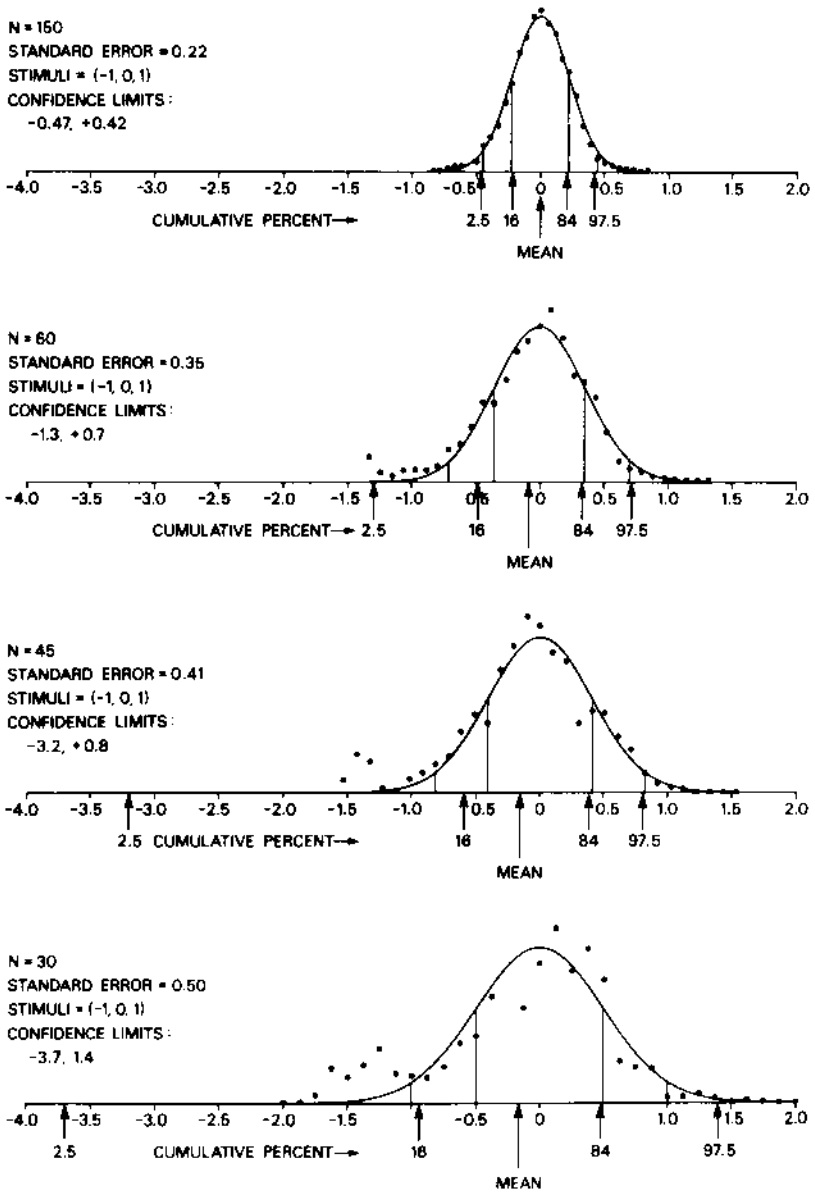

Figure 7. The simulated sampling distributions of the estimated value of T_{75} for the standard case for four different values of N, the total number of trials. Distributions based on 1,200 simulated values using the Monte Carlo technique. The curve drawn through the points is a normal distribution with a standard deviation equal to the calculated standard error (Equation 3) for the indicated sample size. The vertical lines show ±1 and ±2 standard errors. The arrows under the abscissas indicate the stimulus values that correspond to cumulative percentages of simulated values of 2.5%, 16%, 84%, and 97.5%. The arrow labeled "mean" is the average of all simulated values falling between ±4 Z-units.

tle information about the location of the mean and should be avoided, either by increasing the number of stimuli and the number of trials or by constraining the slope's upper limit.[3]

Because of the substantial agreement between the two simulation techniques, we used only the Monte Carlo technique (1,200 simulations) to explore other questions related to sampling strategies. The influence of the total number of trials N on the shape of the sampling distributions of T_{75} is shown in Figure 7 for the standard case. We have plotted the number of simulated values falling within intervals 0.25 Z-units wide for a range extending over ±2 Z-units. The arrows on the abscissas, labeled "Mean," point to the average of all the simulated values of T_{75} falling between ±4 Z-units for each distribution. The curve superimposed on each set of points is a normal

distribution with a standard deviation equal to the standard error calculated from Equation 3 for this sample size. The vertical lines cutting through the normal curves demarcate ±1 and ±2 calculated standard errors. On the abscissa of each graph, arrows have been drawn to show the value that includes ±34% of the simulated distribution and ±47.5% of the stimulated distribution, that is, at the cumulative percentages of the simulated distribution corresponding to 2.5%, 16%, 84%, and 97.5%. If the sampling distribution is normally distributed, the arrows should fall near the vertical lines denoting ±1 and ±2 standard errors, and clearly they do *not* for the three lower distributions.

Three trends were apparent in the simulated sampling distributions as N decreased. First, the width of the distributions increased; second, the confidence limits were no longer equal to twice the calculated standard error; and third, the distributions were asymmetrical. In short, the simulations confirmed our qualitative observations based on the graphical approach. For N=150 trials, the simulated sampling distribution is close to a normal distribution, but for N=60, 45, or 30, the distributions are badly skewed, a skew sufficient to produce a small bias in the mean of the distribution toward the negative side. The observed asymmetry is a property of 2AFC distributions, since the simulated sampling distribution for thresholds estimated from the yes-no psychometric functions is perfectly symmetrical and is well fit by a normal curve even when the total number of trials (N) is as small as 60.

Sampling strategies (Figure 5) can markedly influence both the magnitude of the confidence limits and their asymmetry. In an extended series of simulations of N=60, we manipulated the range (R), number of stimuli (k), and the center of the sampled set ($\bar{X}$). The results of these simulations are plotted in Figure 8. Generally, the simulated confidence interval is larger than ±2 standard errors, which, for 60 trials, should be equal to about ±0.8 Z-units.

Range had a relatively small effect on the *size* of the limits, but sampling with a small range placed a strong constraint on the placement of the stimuli. For the smallest range, small shifts in the position of the stimuli led to a rapid increase in the limits, whereas for the largest range, the limits remained roughly constant for large shifts in the center of the sampled stimulus set. As expected, displacements of the center of the sampled stimulus set toward the high end of the psychometric function were relatively innocuous, whereas equal displacements toward the low end of the function typically produced large variability in the estimated threshold.

For all ranges, the smallest and most symmetrical limits were found with k=2. When a small number of stimuli was used to estimate psychometric functions, the number of trials per stimulus was obviously larger for k=2, with the effect that the intrinsic binomial variability of each

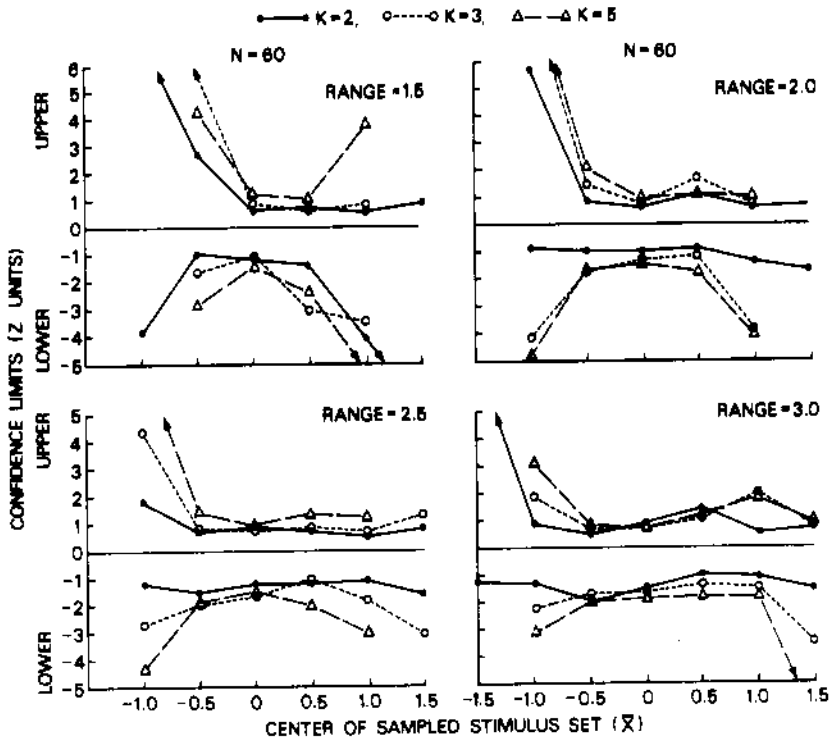

Figure 8. The confidence limits of the estimated value of T_{75} based on Monte Carlo simulations of N=60 for four values of range. As a function of stimulus placement ($\bar{X}$), the continuous line (•) shows the limits for k=2; the dotted line (○) shows the limits for k=3; the dashed line (△) shows the limits for k=5.

sampled point was smaller. Thus, curves with very shallow slopes were less common the smaller the number of tested stimuli. This benefit diminished as the range increased. Because the chance of observing a psychometric function with an infinitely steep slope is substantial for k = 2, a better strategy for small samples is the choice of three to five stimuli and a large range.

Our results from the simulations led to the following conclusions about the statistical properties of 2AFC psychometric functions:

(1) For N < 60, the sampling distribution of the estimate of T_{75} is not normally distributed, but is skewed, usually in a negative direction. The skew is sufficient to introduce a small bias in the estimator.

(2) For small N, the probit analysis equations do not adequately characterize the standard error and the confidence limits associated with the estimate of T_{75}.

(3) For small N, sampling with two stimulus values (k = 2) will produce the smallest confidence interval, but the probability of encountering samples with percents correct less than or equal to 50%, or equal to 100%, is quite high. Since such samples cannot provide an adequate estimate of the slope of the psychometric function, sampling with three stimuli (k = 3) is a better strategy.

(4) For small N, sampling with widely spaced stimuli will generally yield more stable confidence intervals over a broader range of stimulus placements, whereas sampling with closely spaced stimuli will make the choice of stimulus placement critical.

DISCUSSION

The statistical properties of thresholds estimated from 2AFC methods are surprisingly poor, a fact that is not widely recognized. We have argued here that threshold estimates derived from 2AFC techniques are about twice as variable as corresponding estimates derived from yes-no methods and that confidence limits are large, asymmetrical, and not readily predictable from conventional standard error formulas. All of these faults are exacerbated for small samples. The confidence limits for T_{75} in 2AFC experiments do not become well behaved until Ns of 100 or more trials are used. Two questions remain: Are these conclusions inescapable? Are there alternative approaches by which these characteristics can be minimized or avoided?

The Stimulus Domain

The importance of these conclusions obviously depends on the steepness of the psychometric functions expected, the degree of accuracy sought in the estimate of T_{75}, and the number of trials available. Throughout the present paper, the stimuli, standard errors, and confidence limits are scaled in units of σ, the standard deviation of the underlying cumulative normal curve. When psychometric functions are steep in the stimulus domain (σ is small), the standard error and the confidence limits are correspondingly small in that domain, and a given error of estimation of T_{75} in Z-units becomes less serious. On the other hand, as psychometric functions become flatter or the degree of accuracy needed becomes greater in the stimulus domain, the statistical properties of estimates of T_{75} become more important.

Experimental Design

Some practical guidelines for the design of 2AFC experiments follow from the influence of various factors on the magnitude of standard errors and confidence limits. In the typical case, the only factors the experimenter can control exactly are N and k; the range (R) and stimulus placement ($\overline{X}$) are defined with respect to T_{75} and σ. The optimal placement of trials depends on how much information the experimenter brings to the experiment. If values of T_{75} and σ are already rather well known, Figure 8 can be used as a guide to the optimal placement of trials for small samples. If the value of T_{75} is known, but the value of σ is less well known, the range covered by the stimuli should be broadened, because, as Finney (1971, p. 143) says, "a misjudgement causing the actual responses to be a little closer to (T_{75}) than was intended may be catastrophic, whereas responses a little wider apart than intended will usually have less serious consequences."

If T_{75} is not known with any degree of accuracy—if one must necessarily run the risk of a large difference between the center of the sample set ($\overline{X}$) and T_{75}—a larger number of more widely spaced stimulus values must, of course, be used, with the expectation that some of the stimuli will be essentially wasted. What is perhaps less obvious is that when T_{75} is not known with any degree of accuracy, it is better to err in the direction of positive values of $\overline{X}$, that is, in the direction of placing one's stimuli too high rather than too low. The asymmetries of binomial variability lead to the fact that standard errors and confidence intervals change less for positive than for negative shifts of equal magnitude.

Data Selection Rules

At the end of the experiment, the experimenter has a new source of knowledge, the data themselves. In practice, experimenters do not apply statistical analysis blindly to data; rather, they apply ad hoc or explicit criteria to each data set, discarding those data sets that do not yield reasonable estimates of the threshold. For this reason, the simulated distributions of threshold estimates will not be matched in practice.

Discarding of data obviously has disadvantages. In itself, it is not a solution to the problem of small N, because when data sets are discarded, trials are wasted. In addition, unless the criteria for discarding data are formally defined prior to the experiment, discarding data opens the door to experimenter bias. It might be interesting to simulate the use of the 2AFC method of constant

stimuli, generating artificial data sets and using a series of formal criteria for the discarding of data. Some rules for discarding data (e.g., discard data sets that do not conform to a previously designated goodness-of-fit criterion) may create a sampling distribution for small samples with more favorable dispersion properties—smaller standard errors and confidence intervals.[4] It should be possible to seek out and specify data selection rules that provide an optimum benefit/loss ratio, maximizing the goodness of the statistical distribution of estimates, while keeping small the number of worthwhile data sets discarded.

Staircase Methods

The most popular strategy for estimating a threshold from a small sample is the use of staircases or other adaptive procedures (Cornsweet, 1962; Hall, 1981; Watson & Pelli, 1983; Watt & Andrews, 1981). Staircases have one obvious advantage over the method of constant stimuli—the efficient use of trials when little is known about the location of the desired threshold. A well-designed staircase rule will usually place most of the trials in the steeper part of the psychometric function rather than far in the tails. It is sometimes assumed, erroneously in our view, that because staircase estimates of thresholds are efficient, they are always much less variable than any estimate based on the method of constant stimuli. But staircases have no magical power. The accuracy of estimates derived from staircases is constrained by the same factors as is the accuracy of estimation from data collected by the method of constant stimuli—the number of trials, binomial variability, and the shape of the psychometric function. The variability of estimates derived from staircase data *can never be less* than the variability of estimates derived from the method of constant stimuli *selected for the optimal deployment of trials*.

This assertion can be supported on numerical as well as logical grounds. Rose, Teller, and Rendleman (1970) used a computer simulation technique to generate standard errors of estimation for 2AFC staircase. These simulated standard errors can be compared with the standard errors estimated from Equation 3 in the present paper for the 2AFC method of constant stimuli.[5] A selected combination of parameters from the two studies, matched as closely as possible in terms of stimulus spacing (step size) and number of trials, is listed in Table 1. As expected, the standard errors of the staircase estimates sometimes approach, but are never smaller than, the standard errors of estimates from the method of constant stimuli with optimal placement of trials when comparable step sizes and number of trials are used.

Alternative Procedures When N Is Small

We have only a few suggestions for alternative approaches in situations allowing fewer than 100 trials. If one can assume a cumulative normal function with the upper asymptote D = 100%, the threshold criterion can be shifted from 75% to 83%, where the intrinsic error is minimal. This shift is not advisable if the upper asymptote is below 98%, as is common with inattentive subjects. If the tested population is well characterized and the slope of the psychometric function known, it might be reasonable to use a fixed slope approach, where the available data are used to estimate only the location parameter. Any difference between the assumed fixed slope and the true slope of the tested individual may introduce a bias in the estimate, but this cost may be tolerable given the necessary imprecision inherent in the use of small N.

Instead of attempting to estimate a threshold with an inadequate number of trials, one could test each subject on only a single stimulus value. On the basis of prior normative data, that value could be chosen to be high on the psychometric function, perhaps at 83%, in the region where the intrinsic error is smallest for average normal subjects. The distribution of performance levels for normal subjects could be established. An individual subject's or patient's performance below the normal range at that stimulus value would then indicate a deficit with respect to the normal population. An approach similar to this one has been used recently to screen infants for possible visual deficits (Dobson, Teller, Lee, & Wade, 1978; Fulton, Manning, & Dobson, 1978).

Table 1
Comparison of Standard Errors for 2AFC Staircases and 2AFC Method of Constant Stimuli

2AFC Staircases (Rose et al., 1971)				2AFC Method of Constant Stimuli				
Step Size (Z-units)	Number of Intervals	Number of Trials	SE (Z-units)	Step Size (Z-units)	k	Range	Number of Trials	Optimal SE* (Z-units)
0.26	10	50	0.30	0.25	5	1	50	0.33
		200	0.20				200	0.16
0.51	5	50	0.40	0.50	5	2	50	0.37
		200	0.26				200	0.18
0.64	4	50	0.40	0.66	5	2.5	50	0.40
		200	0.20				200	0.20
1.28	2	50	0.66	1.25	3	2.5	51	0.42
		200	0.20				201	0.21
2.56	1	50	0.97	2.50	2	2.5	50	0.49
		200	0.56				200	0.24

**For given values of range and k, the SE given is the minimum value found over all values of $\bar{X}$; that is, the optimum placement of trials is chosen.*

In summary, the statistical properties of forced-choice methods are poorer than might be wished. Although a few alternatives remain to be explored, we believe that this variability is largely unavoidable and will have to be taken into account in the design and interpretation of forced-choice experiments.

We conclude with the reminder that statistical sampling fluctuations provide only one source of the variability in real experiments, and that the minimization of other sources of bias and variability must be combined with statistical theory in the design of experiments to be performed on real subjects.

REFERENCES

CORNSWEET, T. N. (1962). The staircase method in psychophysics. *American Journal of Psychology*, **75**, 485-491.

DOBSON, V., TELLER, D. Y., LEE, C. P., & WADE, B. (1978). A behavioral method for efficient screening of visual acuity in young infants. I. Preliminary laboratory development. *Investigative Ophthalmology and Visual Science*, **17**, 1142-1150.

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

FULTON, A., MANNING, K., & DOBSON, V. (1978). A behavioral method for efficient screening of visual acuity in young infants. II. Clinical application. *Investigative Ophthalmology and Visual Science*, **17**, 1151-1157.

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

ROSE, R., TELLER, D. Y., & RENDLEMAN, P. (1970). Statistical properties of staircase estimates. *Perception & Psychophysics*, **8**, 199-204.

TELLER, D. Y. (1985). Psychophysics of infant vision: Definitions and limitations. In G. Gottlieb & N. Krasnegor (Eds.), *Measurement of audition and vision in the first year of postnatal life: A methodological overview*. Norwood, NJ: Ablex.

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

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

NOTES

1. Qualitatively speaking, the intervals are too large because the vertical bars represent individual confidence intervals for n = 100 trials per point, whereas the confidence limits for T_{75} should reflect the joint probability of obtaining all three data points, with a total N of 300 (see appendix for quantitative formulation).

2. The differences between the two techniques reflected differences in programming strategies, choices usually made to prevent program "crashes" during the long simulation runs. For example, in the enumeration approach, negative or zero slopes were replaced by a very small Slope (0.001) in order to assign the estimated value of T_{75} to the correct tail of the distribution. In the Monte Carlo approach, a special problem arose in assigning a value to percents correct either equal to 100% or equal to or below 50%. Mathematically, these values lay infinitely far from the center of a cumulative normal distribution. In the Monte Carlo simulations, these percentages were initially assigned a value of 4 Z-units. Embedded in the probit analysis program was an iterative routine which oscillated between functions corresponding to these initial values and functions based on much larger values. For these extreme percentages, the program failed to converge, and after six iterations, the value of T_{75} from the last estimated psychometric function was stored, and the program continued on to the next data set.

3. The probability of an infinite slope for k = 2 is high even for N = 60 and stimuli placed near the center of the psychometric function. There is about a 30% chance that one of the two stimuli will fall at .5 or below or at 1.0 if the true percents correct = .58 and .92 (−1 or +1 unit). If k = 3 and the tested stimuli cover the same range (−1, 0, 1), there is only about a 6% chance of an infinite slope.

4. Alternatively, constraints can be placed on the slopes. For example, consider the case of 20 trials at each of the locations Z = −0.5 and Z = +1.5. Table 2 shows that confidence limits estimated by enumeration are −1.9 and +1.5. These enumerations were calculated with the slopes constrained to lie between 0 and ∞. If the upper slope constraint were strengthened to have the slope between 0 and 3, the confidence limits would become −1.9 and 1.2. If the slope were further constrained to be greater than .33, the confidence limits would become −1.35 and 1.2. Thus, a fairly loose constraint on the slope is able to produce a significant reduction in the size of the confidence interval.

5. Since the parameter of range is unspecified in staircase techniques, we have converted the units of both Rose et al. and the present paper to "step size," that is, to the spacing of adjacent stimuli along the stimulus dimension. Conversion of units was carried out as follows. Rose et al. used a ramp psychometric function, and scaled the stimulus axis in terms of the interval $I_o - I_1$ between the lower and upper ends of the ramp. Ancillary simulations showed that ramp and cumulative normal functions yielded quantitatively similar standard errors of estimation. Inspection of Rose et al.'s assumptions and the normal table shows that the interval $I_o - I_1$ corresponds to 2.56 Z-units in our terms. Rose et al. used the number of intervals between $I_o - I_1$ as their parameter of stimulus spacing. That is, if adjacent stimuli were so spaced as to fall at I_o and I_1, the number of intervals between I_o, I_1 is 1 and the step size is 2.56 Z-units; if stimuli fall at I_o, I_1, and halfway between, the number of intervals is 2 and the step size is 1.28 Z-units. In the parameter space of the present paper, the step size is the range divided by the quantity, k − 1; thus, if k = 3 and R = 2, the step size is 1 Z-unit.

APPENDIX

Graphical Analysis

In Figures 2 and 3, a *qualitative* graphical analysis was used to determine confidence limits on the threshold estimate. These

Table 2
95% Confidence Limits

Stimuli	K = 2, R = 2.0, $\bar{X}$ = 0 (−1, 1)		K = 2, R = 2.0, $\bar{X}$ = .5 (−0.5, 1.5)	
N	40	60	40	60
1.96 SE (Equation 3)	−1.01, 1.01	−.83, .83	−.93, .93	−.76, .76
Finney (Equation 4)	indeterminate	−3.0, 0.7	indeterminate	−1.6, 0.7
Enumeration	−1.6, 1.0	−1.0, 1.0	−1.9, 1.5	−1.1, 1.5
Monte Carlo	−1.4, 0.8	−1.0, 0.6	−1.8, 1.3	−0.9, 1.1
Graphical Analysis (Maximum Likelihood)	−1.4, 1.0	−1.0, 1.0	−1.7, 1.5	−1.0, 1.5
Graphical Analysis (chi-square)	−1.2, 1.4	−0.9, 1.0	−1.6, 1.5	−1.0, 1.5

limits were obtained by shifting the position of a straight-line fit in such a way that the threshold was brought to an extreme value, subject to the constraint that the line not lie outside the error bars of any data point. A modification of the constraint based on the chi-square function enables graphical analysis to have *quantitative* validity.

The chi-square function can be written as

$$\chi^2 = \sum_i (P_{Oi} - P_{Ei})^2 / \sigma_{pi}^2, \quad \text{(A1)}$$

where $\sigma_{pi}^2 = P_{Ei}(1 - P_{Ei})/n_i$ and P_E is the underlying expected probability and P_{Oi} is the observed probability. If the difference $\Delta p = P_O - P_E$ is small, then the numerator can be converted to z-scores.

$$\Delta P = \frac{(D-C)}{\sqrt{2\pi}} \exp(-Z^2/2)\Delta Z, \quad \text{(A2)}$$

where D and C are the upper and lower asymptotes. Thus, Equation A1 can be written as

$$\chi^2 = \Sigma (Z_{Oi} - Z_{Ei})^2 / \sigma_{zi}^2, \quad \text{(A3)}$$

where

$$\sigma_{zi}^2 = \frac{2\pi P_{Ei}(1 - P_{Ei})\exp(Z_i^2)}{n_i(D-C)^2}.$$

The weighting function $1/\sigma^2_{zi}$ is precisely the weighting function w_i defined and tabulated by Finney (1971).

The optimal values for threshold and slope would be those values that minimize chi-square. When the threshold or slope deviate from the optimal values, the chi-square will increase. The threshold value that increases chi-square to 1.0 (allowing the slope to adjust to a new optimal value) corresponds to the 68% confidence limit. An increase in chi-square to 4 corresponds to the 95% confidence limit. One of us (Klein) has a probit analysis program available that includes an option for printing out a two-dimensional array of values of chi-square for all values of threshold and slope.

A similar discussion applies when using the likelihood function rather than chi-square, except that a change in the logarithm of the likelihood function is half the corresponding change in chi-square.

In Table 2, we compare the 95% confidence limits for several stimulus locations and for several methods of analysis. The maximum likelihood graphical analysis corresponds quite closely (within 15%) to the exact confidence limits given by the enumerations. The chi-square analysis is almost as good, with significant errors only for the case of 20 trials and stimuli at +1 and −1 Z-units. Finney's formula, given by Equation 4, has a large disagreement with the enumerations, as does +1.96 SE. Thus, maximum likelihood graphical analysis is a highly satisfactory method for estimating confidence limits.

(Manuscript received October 29, 1982;
revision accepted for publication February 25, 1985.)