Perception & Psychophysics
1992, 51 (5), 409-422

Extracting thresholds from noisy psychophysical data

WILLIAM H. SWANSON and EILEEN E. BIRCH
Retina Foundation of the Southwest, Dallas, Texas
and The University of Texas Southwestern Medical Center, Dallas, Texas

Psychophysical studies with infants or with patients often are unable to use pilot data, training, or large numbers of trials. To evaluate threshold estimates under these conditions, computer simulations of experiments with small numbers of trials were performed by using psychometric functions based on a model of two types of noise: *stimulus-related noise* (affecting slope) and *extraneous noise* (affecting upper asymptote). Threshold estimates were biased and imprecise when extraneous noise was high, as were the estimates of extraneous noise. Strategies were developed for rejecting data sets as too noisy for unbiased and precise threshold estimation; these strategies were most successful when extraneous noise was low for most of the data sets. An analysis of 1,026 data sets from visual function tests of infants and toddlers showed that extraneous noise is often considerable, that experimental paradigms can be developed that minimize extraneous noise, and that data analysis that does not consider the effects of extraneous noise may underestimate test-retest reliability and overestimate interocular differences.

Psychophysical measurements of threshold frequently involve large numbers of trials under a range of stimulus conditions and data analysis that defines threshold as a point on a psychometric function. Data analysis typically consists of fitting the data with a psychometric function by using maximum likelihood estimation. The bias and precision of threshold estimates have been shown to depend on the slope of the psychometric function, on the placement of stimuli relative to threshold, and on the total number of trials (McKee, Klein, & Teller, 1985; O'Regan & Humbert, 1989; Rose, Teller, & Rendleman, 1970; Watson & Fitzhugh, 1990).

With normal adults, relatively unbiased and precise threshold estimates can be obtained by training, by using pilot data to guide stimulus placement, and by gathering large numbers of trials. Unfortunately, these strategies are at best difficult and at worst impossible to employ with infants or with patients in a clinical research setting. In these situations, data sets are limited to small numbers of trials and pilot data may not be available to guide stimulus placement. *Stimulus-related noise*, because of variability in the stimulus and in the responses of the visual system to it, may be increased by immaturity and/or disease (i.e., the slope of the psychometric function may be shallower than it is for normal adults; Brown, Dobson, & Maier, 1987; Mayer & Dobson, 1982). In addition, *extraneous noise*, such as responses to irrelevant aspects of the experimental situation, may result in upper asymptotes of less than 100%, increasing bias and decreasing precision of threshold estimates (Green, 1990; Hall, 1981; Klein & Manny, 1989; Madigan & Williams, 1987; Manny & Klein, 1985; McKee et al., 1985; Pelli, Robson, & Wilkins, 1988; Teller, Mar, & Preston, in press).

The only detailed study of the effects of extraneous noise on clinical threshold estimates has been an evaluation of a descending method of limits in a visual contrast sensitivity test (Pelli et al., 1988). However, in a number of clinical and infant tests, two-alternative forced choice (2AFC) is used with either a constant stimuli protocol with widely spaced ($\geq$ 1 octave) stimuli or an adaptive staircase protocol in order to achieve adequate stimulus placement with a limited number of trials and no a priori information about threshold. The present study therefore focused on 2AFC threshold estimates under the "worst case" conditions of small numbers of trials and high amounts of both stimulus-related and extraneous noise. The three aims were to quantify the bias and precision of threshold estimates under these conditions, to develop strategies for rejecting data sets as too noisy, and to evaluate the effects of extraneous noise on forced-choice-preferential-looking (FPL) data. For the first two aims, maximum likelihood estimation was used to analyze 2AFC experiments simulated with Monte Carlo techniques in which psychometric functions having shallow slopes and varying upper asymptotes were used. For the third aim, maximum likelihood estimation was used to analyze 1,048 FPL data sets obtained from visual function tests of 320

This project was supported in part by grants from the National Institutes of Health (EY05236 & EY07716). The project was started after discussions with both Davida Y. Teller and Stanley A. Klein, who also made helpful comments during the course of the research. Andrew B. Watson provided suggestions concerning simulations of maximum-likelihood adaptive methods, and John M. Foley provided suggestions concerning the ROC analysis. Correspondence should be addressed to William H. Swanson, Retina Foundation of the Southwest, 9900 N. Central Expressway, Suite 400, Dallas, TX 75231.

healthy infants and toddlers (data from Birch, 1985; Birch & Hale, 1988; Swanson & Birch, 1990).

METHODS

Model

In a typical 2AFC experiment, a stimulus is presented on each trial in one of two locations and the subject is required to choose which of the two locations contains the stimulus. The stimulus level is varied from trial to trial along the physical dimension of interest (such as intensity, spatial frequency, or contrast). If the range of stimulus levels is chosen appropriately, if the number of trials is sufficiently large, and if the subject is fully cooperative, the percentage of correct responses will decrease monotonically from 100% to 50%. Such data sets are well described by a function, $R(x)$, which gives the probability of a correct response for 2AFC with stimulus level x:

$$R(x) = P(x) + 0.5[1-P(x)], \quad (1)$$

in which $P(x)$ gives the probability of detecting the stimulus. Quick's (1974) version of the Weibull function was used to define $P(x)$:

$$P(x) = 1 - 2^{-(x/\alpha)^{\beta}}, \quad (2)$$

in which α is threshold and β controls the slope.

For a sufficiently large number of trials, the subject will be correct on the fraction $P(x)$ of the trials because the stimulus was seen, and will be correct by chance on half of the remaining fraction $[1-P(x)]$ of the trials. If the subject is fully cooperative but the range of stimulus levels is not appropriate and/or the number of trials is too small, the function may not be monotonic and may not reach either 100% or 50% correct, but Equation 1 can still be used to describe the data.

If the subject responds to irrelevant aspects of the experimental situation or mistakenly makes the wrong response even though the stimulus was detected, $R(x)$ may never reach 100% correct, even if the range of stimulus levels is appropriate and the number of trials is large. To model the effects of extraneous noise, Equation 1 was modified:

$$R(x) = \gamma P(x) + 0.5[1-P(x)], \quad (3)$$

in which γ is the upper asymptote of $R(x)$.[1] Figure 1 shows examples of $R(x)$ for several values of β (upper panel) and γ (lower panel).

The purpose of the current paper is to explore the consequences of adding one additional component (extraneous noise) to the commonly used model of a single source of noise (stimulus-related noise). In fact, other forms of noise could also be expected. For example, if the subject becomes fatigued or habituated as the experiment proceeds, the amount of noise could increase. On the other hand, noise could decrease during the course of the experiment because of practice effects. Such sources of noise are certainly worth consideration but are outside the scope of the current study.

Parameter Estimation

Maximum likelihood estimation (Harvey, 1986; Watson, 1979) was used to fit each data set with Equation 3. For a given stimulus level x, the likelihood $L(x)$ that correct responses are obtained on k out of n trials is given by

$$L(x) = [n!/(k!(n-k)!)]\,[R(x)]^k\,[1-R(x)]^{n-k}, \quad (4)$$

and the likelihood of a complete data set for a given experiment is the product of the likelihoods for all of the stimulus levels used in that experiment. Preliminary simulations showed that, with three parameters, searching algorithms tended to get caught in local minima. To avoid this problem, likelihoods were computed for a fixed range of parameters (α,β,γ), and the parameter set that gave the maximum likelihood for an individual data set yielded the estimates. The parameter sets evaluated varied γ from 76% to 100% in steps of 2% and β from 0.8 to 14.2 in steps of 0.25 log unit. For the simulations, α was varied from 0.0 to 4.0 in steps of 0.1 log unit; for the fitted data, α was varied in steps of 0.1 log unit across the range of available stimulus levels.

To compare maximum likelihood threshold estimation with more common methods of estimating threshold, threshold estimates were also generated with other techniques. For the simulations of staircase experiments, comparisons were made with the means of reversals of the staircases.[2] For the analysis of published data, comparisons were made with values from the original studies, in which the constant stimuli data were analyzed in a graphical manner (Birch, 1985, estimated the 75% correct point for constant stimulus data by interpolating between points on the psychometric function) and the staircase data were analyzed by taking the mean of all but the first two reversals.

Simulations

To determine the effects of extraneous noise on threshold estimation under the difficult conditions of a small number of trials, a shallow slope for $P(x)$, and upper asymptotes less than 100%, Monte Carlo simulations of 2AFC experiments limited to 20-60 trials[3] were performed by using psychometric functions $P(x)$ with $\beta = 2$ and γ ranging from 85% to 100%. Both constant stimuli and staircase experiments were simulated for a range of stimulus distributions.

Monte Carlo simulations were performed as in previous studies (e.g., Madigan & Williams, 1987; McKee et al., 1985; O'Regan & Humbert, 1989; Watson & Fitzhugh, 1990). For a given parameter set (α,β,γ), experiments were simulated by using Equation 3. The simulated data were analyzed by using maximum likelihood estimation, resulting in an estimated parameter set $(\alpha_{est},\beta_{est},\gamma_{est})$. For a given parameter set (α,β,γ), means and standard deviations for α_{est}, β_{est}, and γ_{est} were computed from the results of 100-1,000 simulations. *Bias* of parameter estimates was calculated as the difference between the actual parameter value and the mean of the estimates. *Precision* of the parameter estimates was calculated as the standard deviation of the estimates. For both bias and precision, calculations were in log units for α and β and in linear units for γ.

To avoid the need to refer to a particular physical dimension of the stimulus, stimulus intensity was expressed relative to the maximum stimulus level employed. Log threshold (log α) was set to 0.0 when threshold was equal to the maximum stimulus level, and simulations varied log α in 0.1-log-unit steps from −3.0 to 0.0.

For the constant stimuli simulations, 1,000 simulations were run for each parameter set (α,β,γ). Optimal stimulus distributions were determined by simulating constant stimuli data for 60 trials, with γ ranging from 100% (no extraneous noise) to 85% (extraneous noise affecting 30% of trials) in steps of 5%. Five stimulus intensities were used, ranging in 1.0-octave steps from the maximum intensity to 1.2 log units below maximum, with 12 trials/intensity.

For the staircase simulations, a staircase protocol in wide use was simulated; this chooses the stimulus level for each trial by a 2-down-1-up decision rule, with a step size of 0.5 octave. All staircases began at maximum stimulus intensity and continued until 10 reversals were obtained; variations in this protocol were also explored, as discussed in the section titled "Staircases." One hundred simulations were run for each parameter set (α,β,γ).

RESULTS

Simulation

Threshold Estimates

The results of 1,000 constant stimulus simulations per value of α, for $\gamma = 100\%$, are shown in Figure 2. Threshold estimates were relatively unbiased (log α_{est} − log α near 0.0) and precise (standard deviation of log γ_{est} near

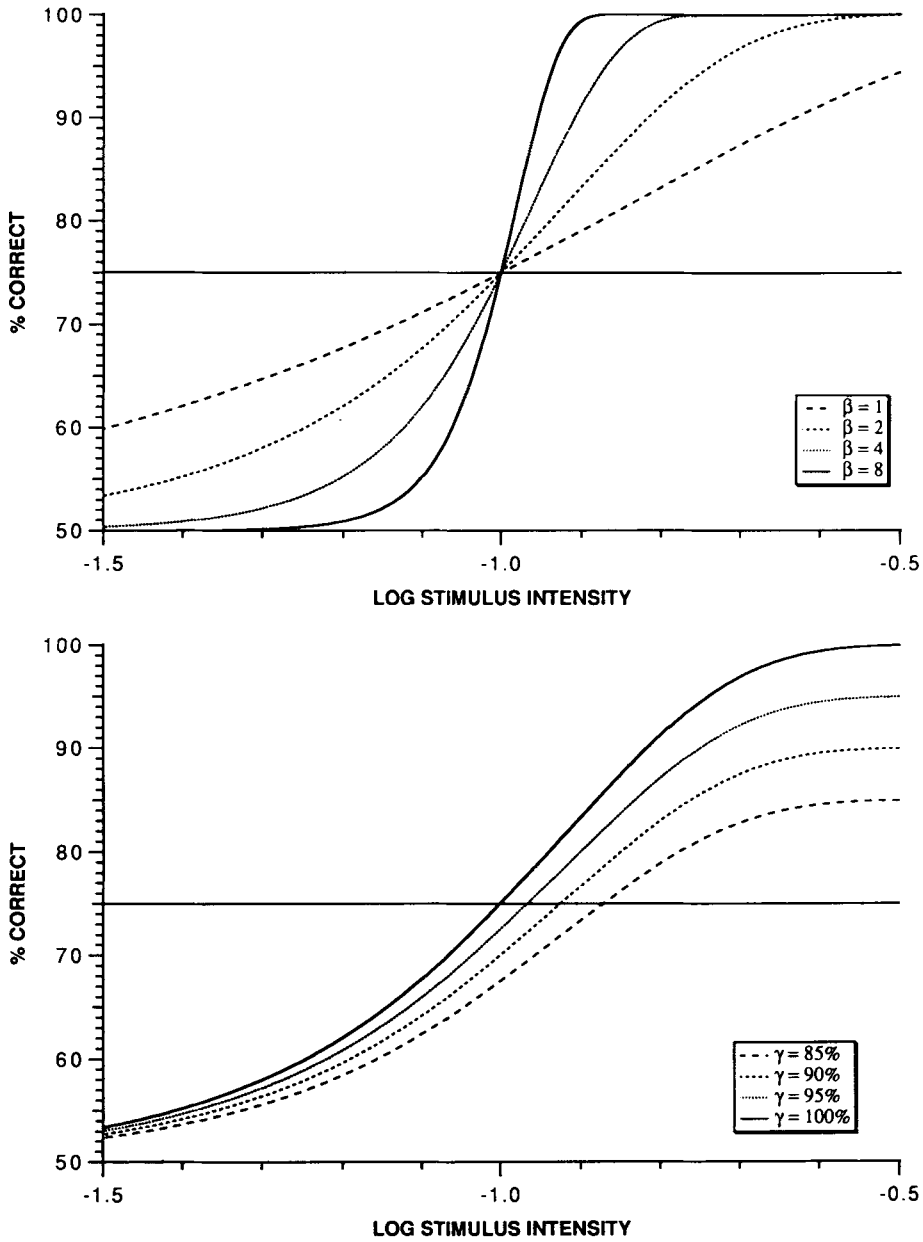

Figure 1. Theoretical psychometric functions, $R(x)$, calculated from Equation 3 for several values of β with γ = 100% (upper panel) and for several values of γ with β = 2.0 (lower panel). Since β is inversely proportional to stimulus-related noise, $P(x)$ becomes steeper as β increases. Note that $P(x)$ also becomes steeper as γ increases.

0.0) only when log α (the actual threshold) was near the middle of the stimulus range. Since there is a ceiling effect when the actual threshold is near 0.0, it may appear that the situation is not so bad for high thresholds; however, if the simulations ignore the ceiling and allow log α_{est} to exceed 0.0, the standard deviations increase dramatically in this region. For $\gamma < 100\%$, means were similar, but standard deviations were up to twice as large. This example illustrates the problem of not having an a priori threshold estimate. For experiments with only a limited number of trials, the method of constant stimuli is therefore only appropriate when a very good a priori threshold estimate can be used to guide stimulus placement.

Several additional constant stimuli simulations were performed to determine whether different choices for the stimulus distribution could provide unbiased and precise threshold estimates with less stringent demands on the accuracy of the a priori threshold estimate. When the step size between stimulus levels was increased to 1.5 or 2.0 octaves, unbiased thresholds were recovered over a wider range for log α, but precision decreased because of the fact that fewer stimuli were near threshold. When the step

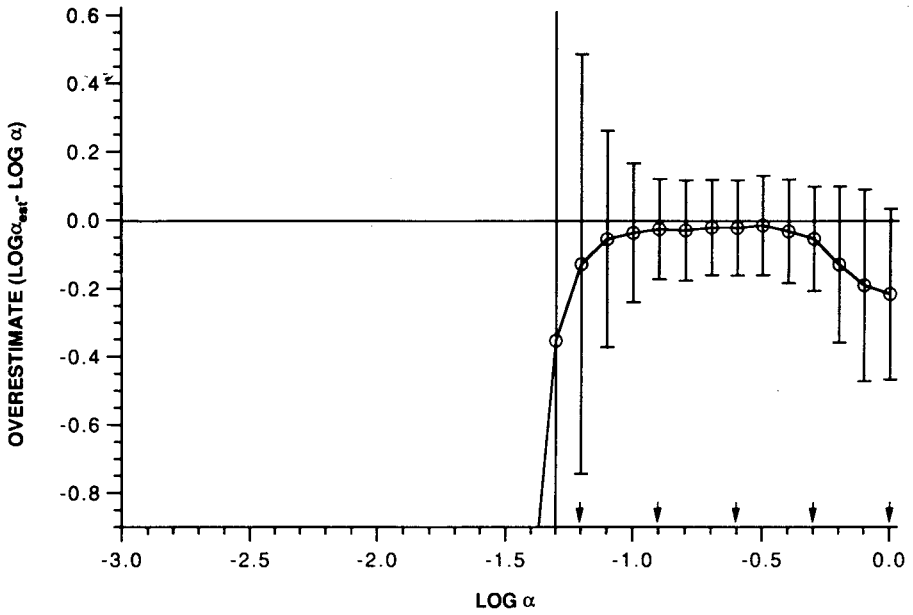

Figure 2. Bias (mean of log α_{est} − log α) and precision (standard deviation of log α_{est}) of threshold estimates for the method of constant stimuli for 1,000 simulations/condition with γ = 100%. There were 12 trials each at five stimulus levels, indicated by arrows on the *x*-axis.

size between stimulus levels was decreased to 0.5 octave, with six trials at 10 stimulus levels (in order to give the same total number of stimuli and the same range of stimulus intensities), results were almost identical to those in Figure 2. Therefore, precision of threshold estimates for the method of constant stimuli could not be improved by changing the step size.

Results of maximum likelihood estimation on the stimulus distributions obtained from staircases are shown in Figure 3, for 100 staircase simulations per condition. When there was no extraneous noise (γ = 100%), bias was low over most of the three-log-unit range of threshold values (for log $\alpha > -0.3$, the estimates were a little low; this is a ceiling effect), and precision was good (standard deviations were generally less than one step size). This performance is comparable to that for data gathered with the method of constant stimuli in which an optimal stimulus placement is used, indicating that when there is no extraneous noise staircases can provide optimal stimulus placement without requiring an a priori estimate of threshold.

High levels of extraneous noise counter the benefits of staircases. When extraneous noise was moderate (γ = 95%), precision remained good, and bias increased only slightly for values of log $\alpha > -2.0$. When extraneous noise was high (γ = 90% or 85%), decreases in log α were accompanied by dramatic increases in bias and decreases in precision, with thresholds tending to be overestimated.

Stimulus-related Noise Estimates

For both the constant stimuli and staircase simulations, the amount of stimulus-related noise tended to be underestimated ($\beta_{est} > \beta$), and the precision was poor (standard deviation was 0.3 to 0.4 log unit, independent of α). This means that even if maximum likelihood estimation suggests that stimulus-related noise is low (i.e., β_{est} is large), it is still possible that stimulus-related noise was in fact high.

Extraneous Noise Estimates

Estimates of extraneous noise for staircase data are shown in Figure 4. When γ = 100% and log $\alpha < -0.5$, the estimates of extraneous noise tended to have little bias and good precision but, as γ decreased, the bias tended to increase and the precision tended to decrease. In practical terms, this means that when the mean γ_{est} for group data is near 100% with a small standard deviation, then most individual data sets must have had minimal extraneous noise. On the other hand, γ_{est} near 100% for an individual data set does not indicate unequivocally that extraneous noise was minimal for that data set.

An additional source of information, the logarithmic mean of stimulus intensities at the staircase reversals, can be used to estimate the amount of extraneous noise for a given data set. The difference between log α_{est} and the mean of reversals was computed for each simulated staircase; the means and standard deviations for these differences are shown in Figure 5. For γ = 100%, the two estimates were quite similar; for $\gamma < 100\%$, the means and standard deviations of the difference increased as log α decreased. In general, for $\gamma < 100\%$, the maximum likelihood estimates of threshold were considerably smaller (and hence closer to the actual threshold) than the mean-of-reversals estimates of threshold. This suggests that when an individual data set yields similar threshold esti-

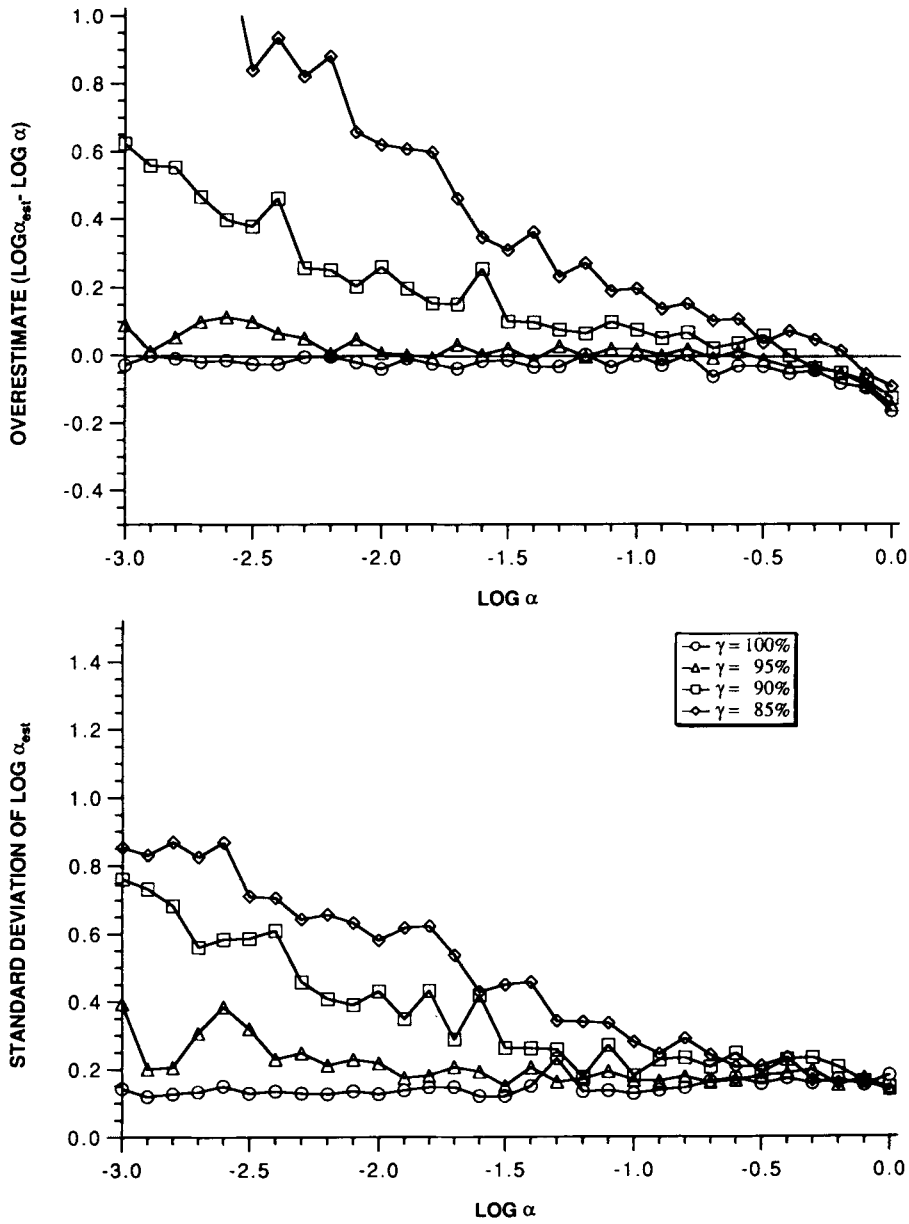

Figure 3. Bias (mean of log $\alpha_{est} - \log \alpha$) and precision (standard deviation of log α_{est}) of threshold estimates for staircases, with 0.5 octave steps, for 100 simulations/condition with γ ranging from 85% to 100%.

mates with both mean of reversals and maximum likelihood estimation, extraneous noise probably had little effect on the threshold estimate.

The primary reason that extraneous noise had a greater effect on mean of reversals than on the maximum likelihood estimate is that because of extraneous noise, errors can occur early in the staircase, at stimulus levels well above threshold. The first error will be the first reversal, followed by a second reversal with the subsequent correct responses. Each error caused by extraneous noise will therefore contribute two reversals near that stimulus intensity, which may be well above threshold. Some researchers have attempted to compensate for effects of extraneous noise on staircase estimates by excluding early reversals in the mean of the reversals. The rationale for this approach is that incorrect responses that occur early in the staircase may have a starting point bias (e.g., Nachmias, 1982). Simulations showed that means of reversals did indeed yield lower threshold estimates when the first two reversals were excluded, with the greatest effects for the smallest values of log α. For γ = 100%, the decrease in threshold was always less than 0.03 log unit; for γ = 95%, the decrease in threshold was as much as 0.26 log unit. The decrease in threshold was smaller for lower values of γ: no more than 0.23 log unit for γ = 90%, and no more than 0.15 log unit for γ = 85%. Ex-

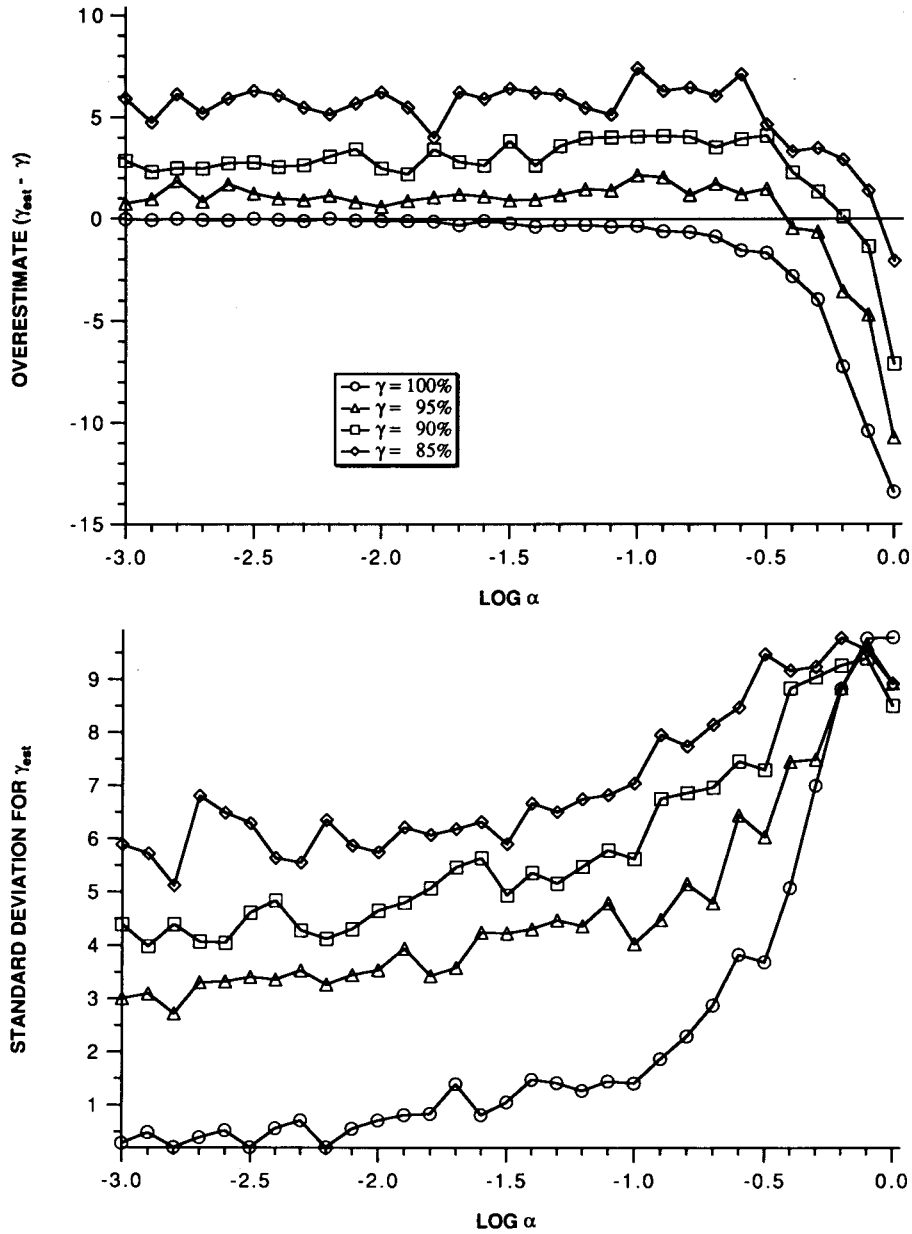

Figure 4. Bias (mean of $\gamma_{est}-\gamma$) and precision (standard deviation of γ_{est}) of the extraneous noise estimates for the staircases for 100 simulations/condition with γ ranging from 85% to 100%.

cluding even more reversals resulted in even greater improvements for high amounts of extraneous noise. The best case was for simulations run with a step size of 1 octave for the first 4 reversals and 0.5 octave thereafter. When 12 reversals were obtained, and only the last 4 reversals averaged for the threshold estimate, the difference between the mean of reversals and the maximum likelihood threshold estimate was quite small, even for γ = 85%. However, when extraneous noise was high, both threshold estimates were still biased and imprecise. Since a discrepancy between the two estimates may indicate that threshold estimates are biased, it may not be desirable to reduce the difference between the two estimates.

One approach to decreasing bias and improving precision of γ_{est} is to include *free trials* throughout the staircase; these are trials at the maximum stimulus level, for which responses do not drive the staircase but can be used in maximum likelihood estimation. Simulations were run in which the number of free trials ranged from 10% to 30% of the number of trials in the staircase, and two different types of data analysis were used. First, γ_{est} was determined by allowing all three parameters (α_{est}, β_{est}, γ_{est}) to vary. Second, γ_{est} was set equal to the percentage of correct responses to the maximum stimulus intensity, and the data were fitted to allow only α_{est} and β_{est} to vary. With the number of free trials equal to 30% of the total

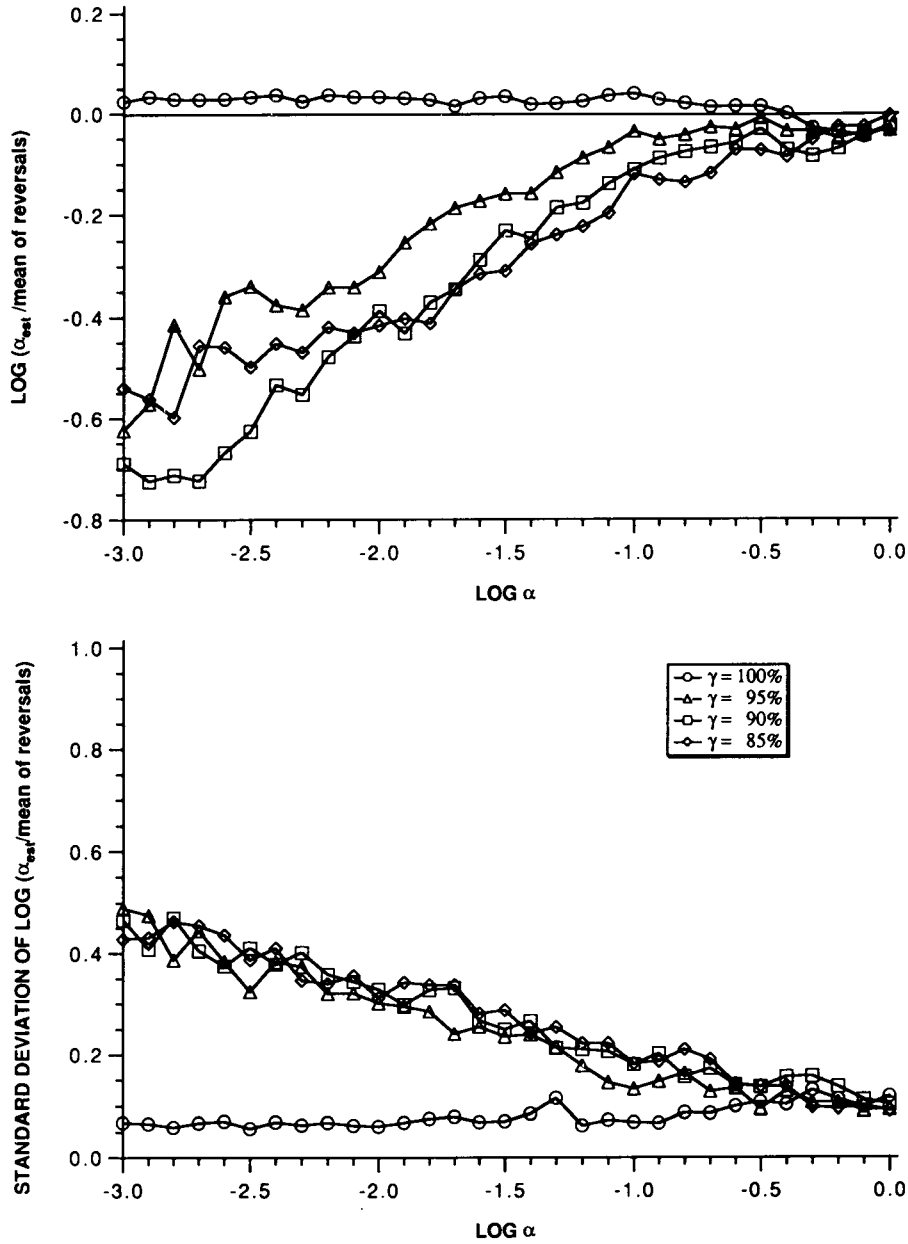

Figure 5. Comparison of maximum likelihood and mean-of-reversals threshold estimates. The upper panel shows the mean difference between the two estimates as a function of log α; the lower panel shows the standard deviation of the difference as a function of log α.

number of trials, the first strategy reduced the bias for γ_{est} by as much as a factor of 2, and the second strategy reduced the bias by as much as a factor of 6; in neither case was the precision substantially affected.[4] Therefore, including free trials can decrease the bias of γ_{est} but cannot improve its precision.

Criteria for Rejecting Individual Data Sets

Based on the preceding analysis of bias and precision of parameter estimates, several strategies were explored for establishing criteria used to reject individual data sets as unreliable. Simulations were performed for each of three hypothetical populations, with distributions of values for γ as shown in the lower portion of Figure 6: (1) high amounts of extraneous noise for most data sets (left), (2) amount of extraneous noise uniformly distributed (middle), and (3) low amounts of extraneous noise for most data sets (right). For all staircases, free trials were included as a fraction (30%) of the total number of trials generated by the staircase. For each population, the staircases for a given value of γ were divided into seven equal groups on the basis of log α, ranging from −3.0 to 0 in 0.5-log-unit steps. A total of 1,008 simulated data sets were generated for each of the three populations. For each

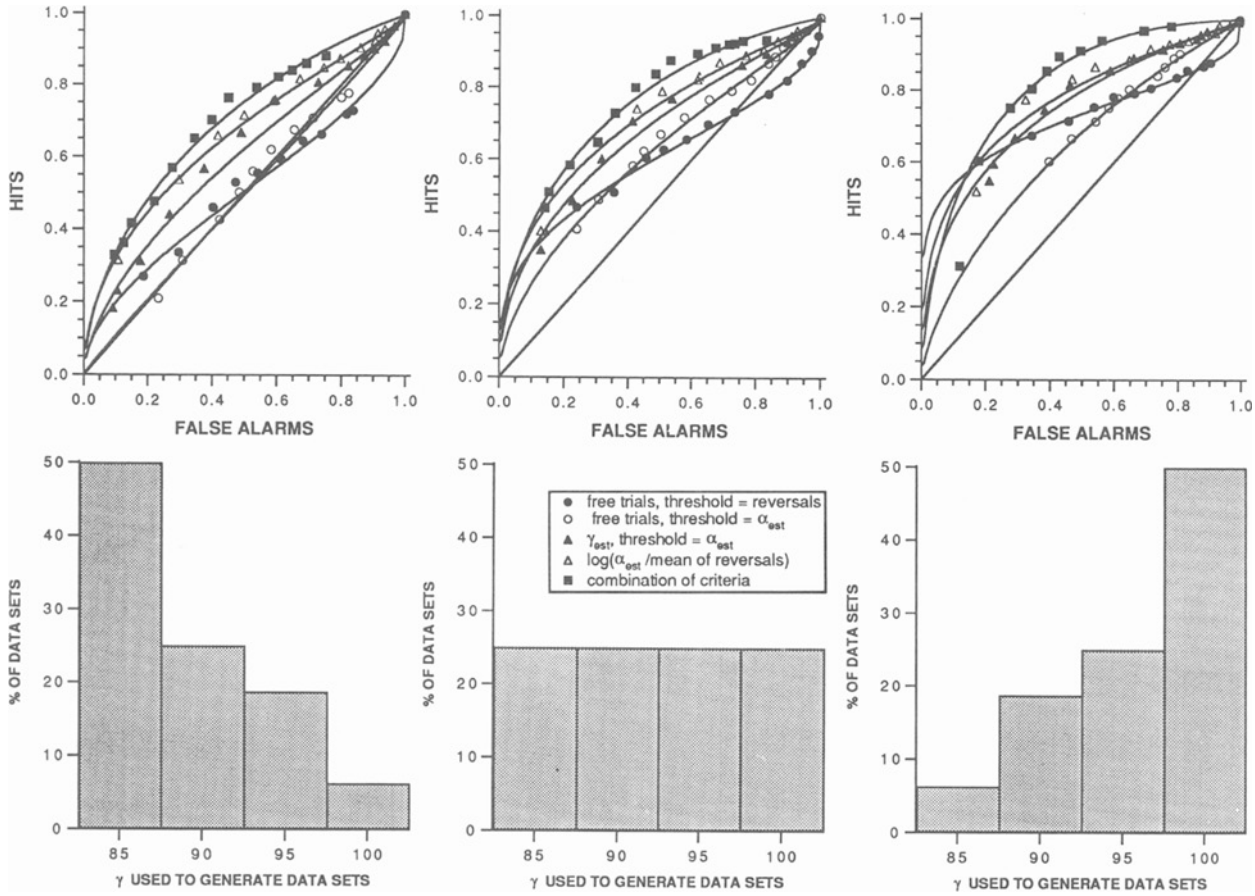

Figure 6. Receiver-operating characteristic (ROC) curves (upper) from five strategies for rejecting individual data sets, based on simulations of populations with three distributions of γ (lower): high amounts of extraneous noise in most data sets (left), amount of extraneous noise uniformly distributed (middle), and low amounts of extraneous noise in most data sets (right). The diagonal lines in the ROC plots indicate performance at the level of chance. Solid circles represent results for the first strategy, open circles for the second strategy, solid triangles for the third strategy, open triangles for the fourth strategy, and solid squares for the fifth strategy. See text and Table 1 for details.

combination (γ,α), the number of simulated experiments was between 4 and 72, as determined by the percentage of data sets for a given population having that value of γ. The simulated data sets were evaluated with five different strategies, which are outlined in Table 1. The strategies differed in the method for obtaining γ_{est}, the method for obtaining the threshold estimate, and the criterion for rejecting data sets as unreliable.

For each population and strategy, the success of the strategy was evaluated with receiver-operating-characteristic (ROC) methodology. ROC methodology was first applied to statistical decision theory by radiologists (Lusted, 1967) and has since been used for a wide range of decision processes (reviewed by Metz, 1982). For the current application, the decision to be evaluated was whether or not a given data set should be rejected as inadequate for reliable threshold estimation. For the ROC analysis, a *true positive* was a data set accepted by the criterion for which the threshold estimate was within 0.2 log unit of the actual threshold (log α), and a *false positive* was a data set accepted by the criterion for which the threshold estimate was not within 0.2 log unit of log α. For a given population and strategy, the fraction of true positives was plotted as a function of false positives for a range of criteria, generating[5] an ROC curve. The area under the ROC curve was used as an index of the success of the strategy: Performance at chance level yields ROC area = 0.5; perfect discrimination yields ROC area = 1.0.

For example, in the first strategy, we used the mean of reversals as the threshold estimate and rejected data sets for which the fraction of correct responses to free trials was less than a specified lower limit. Thirteen lower limits from 76% to 100% (steps of 2%) were evaluated and, for each lower limit, the true-positive fraction was plotted against the false-positive fraction. For Population 1, with high amounts of extraneous noise for most data sets, ROC area = 0.50. For Population 3, with low amounts of extraneous noise, ROC area = 0.72.

ROC results for five strategies are shown in the top portion of Figure 6. The straight diagonal lines show performance at the level of chance. The first strategy (in which we used the mean of reversals for the threshold

Table 1
Strategies for Rejecting Data Sets as Unreliable, Used to Construct ROC Curves in Figure 6

Strategy	1	2	3	4	5
"True Positive" defined as threshold estimate within 0.2 log unit of log α using:					
log mean of reversals	•				
log α_{est}		•	•	•	•
Reject data set if:					
% of correct responses to free trials < criterion	•	•			
γ_{est} < criterion			•		•
log α_{est} − log mean of reversals < criterion				•	•
Regardless of other rejection criteria, accept data set if:					
log α_{est} < −2.9 or log α_{est} > −0.2					•

Note—For the second strategy, log α_{est} was determined with γ_{est} fixed equal to the fraction of correct responses to free trials, whereas for the Strategies 3, 4, and 5, log α_{est} was determined with γ_{est} allowed to vary.

estimate) was always less successful than the use of information from maximum likelihood estimation and was no better than chance for the population with high amounts of extraneous noise for most data sets. The fifth strategy (a combination of criteria) was always the most successful (ROC area = 0.81 for Population 3, 0.75 for Population 2, and 0.70 for Population 1). This analysis shows that the optimal strategy for obtaining unbiased threshold estimates is to design a test paradigm that minimizes the number of data sets affected by extraneous noise and to utilize information obtained with maximum likelihood estimation.

Extraneous Noise and Stimulus Distributions

Staircases may not yield optimal stimulus distributions when extraneous noise is high, since incorrect responses caused by extraneous noise can prevent the staircase from reaching sufficiently low stimulus intensities, as illustrated in Figure 7. When both the upper asymptote and the threshold are low, only a small fraction of trials fall within 1 octave of threshold.

Several strategies to improve staircase stimulus placement were evaluated. Three kinds of staircases were simulated: constant step size throughout (1-octave or 0.5-octave steps), 1-octave steps to the fourth reversal and 0.5 octave steps thereafter, or 1-octave steps until the first reversal and then a reduction in step size with each reversal until a minimum step size of 0.5 or 0.25 octave was obtained. All three gave similar results to those shown in Figure 7; in some cases, the bias was reduced slightly, but in no case did the precision improve. To obtain more stimuli near threshold, after every 5 presentations a stim-

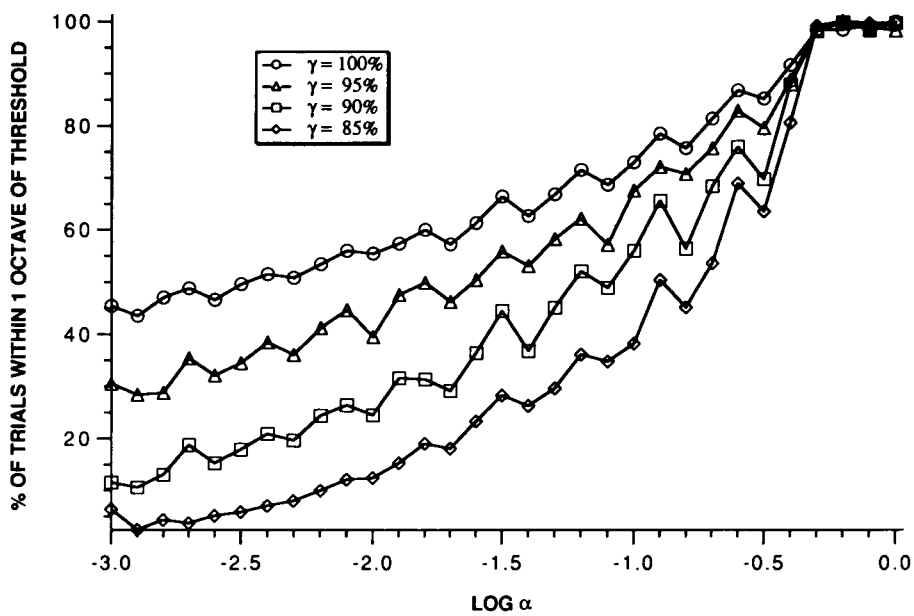

Figure 7. Effects of extraneous noise on stimulus distributions obtained with a 2-down-1-up staircase. The percentage of trials within 1 octave of threshold is shown as a function of log α.

ulus was presented that was 1-2 octaves less intense than the current staircase level; the responses did not drive the staircase but were included in the maximum likelihood analysis. This approach decreased the mean values for log α_{est} − log α but doubled the standard deviation of the estimate.

An alternative to staircase procedures is to use more complicated methods employing ongoing maximum likelihood estimates of the psychometric function to guide stimulus placement (e.g., Harvey, 1986; Madigan & Williams, 1987; Pentland, 1980; Taylor & Creelman, 1967; Watson & Pelli, 1983). These assume fixed values of γ_{est} and β_{est}, varying only α_{est}, and use an a priori likelihood distribution to guide initial trials in the staircase. To evaluate the effects of γ that is lower than 99%, simulations were performed assuming β_{est} = 2.0 and γ_{est} fixed at values ranging from 85% to 99%. No fixed value for γ_{est} yielded unbiased and precise threshold estimates for all levels of extraneous noise. When γ_{est} was fixed at 95%, results were similar to those for staircases. Fixing γ_{est} at lower values resulted both in lower precision in threshold estimates than for staircases and in overestimates of threshold when γ was near 100%. These deficiencies stem from the fact that stimulus placement is inadequate when the actual amount of extraneous noise differs from the fixed value of γ_{est}.

Comment

The simulations showed that constant stimulus methods require optimal stimulus distributions for unbiased and precise threshold estimates. Watson and Fitzhugh (1990) came to a similar conclusion by using a somewhat different approach. If there is no a priori threshold estimate, staircase methods are more suitable for obtaining adequate stimulus distributions. If extraneous noise is low, staircase estimates can be as unbiased and precise as optimal constant stimulus experiments; if extraneous noise is high, then staircases yield stimulus distributions nearly as unsuitable as those from constant stimulus experiments. Since it is difficult to estimate the amount of extraneous noise for individual data sets when the number of trials is low, the best strategy is to develop experimental designs that minimize extraneous noise (as estimated from group data) and to utilize information from maximum likelihood estimation in developing criteria for rejecting individual data sets.

FPL DATA

Database

Four groups of forced-choice-preferential-looking (FPL) data sets gathered from normal infants and toddlers were fitted with Equation 3. One group was composed of 294 monocular and binocular grating acuity data sets gathered from 0-11-month-old infants with the method of constant stimuli (Birch, 1985). These experiments used 12 trials for each of five spatial frequencies, which were spaced in octave steps. Stimulus placement was based on expected age norms. Fourteen infants had at least 1 data set for which the estimated acuity (α_{est}) was higher than the highest spatial frequency used, so the 36 data sets from these 14 infants were excluded from the subsequent analyses. A second group was composed of 312 grating acuity data sets gathered from 0-11-month-old infants with a staircase method (Birch & Hale, 1988). Each staircase started with a 0.38 c/deg grating and proceeded to higher spatial frequencies with a 2-down-1-up decision rule, with approximately 0.5-octave steps and a total of 10 reversals. At random intervals during the course of the staircase, low spatial frequencies (0.38 and 0.75 c/deg) were presented as free trials to maintain interest. A third group was composed of 351 grating acuity data sets gathered from 17-61-month-old toddlers with the same staircase procedure, using an operant technique that involved training with 0.38 c/deg gratings prior to the start of the staircase and a food reward for each correct response. A fourth group was composed of 128 contrast-sensitivity data sets gathered from 4-8-month-old infants with a staircase method (Swanson & Birch, 1990). Each staircase started with 100% contrast and proceeded to lower contrasts with a 2-down-1-up decision rule, with a step size of 1.0 octave until the first reversal, then 0.5 octave until the second reversal, and then 0.25 octave until a total of 8 reversals were obtained. When the running mean of reversals was within 1 octave of 100% contrast, an alternate block method (Dobson, 1983) was used. For 23 of the data sets the block method was used and the fraction correct at 100% contrast was less than 75%, so these data sets were excluded from the analysis.

Extraneous Noise for Different Experimental Methods

The extraneous noise estimates for the four groups are given in Table 2. These results indicate that extraneous

Table 2
Extraneous Noise Estimates for Different Experimental Methods Used With Normal Infants and Toddlers

Method	No. of Data Sets	Mean γ_{est} (in %)	% of Data Sets With: γ_{est} > 95%	γ_{est} < 85%
Constant stimuli (infant acuity)	258	93.1±7.8	53.5	19.8
Staircase (infant acuity)	312	90.6±7.0	29.2	20.9
Operant staircase (toddler acuity)	351	99.8±0.1	99.7	0
Staircase(infant contrast sensitivity)	105	97.1±6.1	77.1	8.6

noise can be significant for data sets gathered from infants. For two of the groups (infant constant stimuli and staircase acuity data sets), the mean value of γ_{est} was near 90%. In comparison, the operant staircase toddler acuity data sets and the infant contrast-sensitivity staircase data sets had fairly high values for γ_{est}. This indicates that extraneous noise can be significant for some data sets and that experimental paradigms can be devised that greatly reduce the amount of extraneous noise for most data sets gathered from inexperienced subjects.

The simulations showed that the bias of maximum likelihood threshold estimates was slight in the presence of extraneous noise for constant stimuli data sets with appropriate stimulus distributions but, for staircase data sets, bias could be significant in the presence of extraneous noise. Since the simulations also showed that γ_{est} tends to be an overestimate when γ is low, it is possible that most of the infant staircase acuity data sets had $\gamma < 90\%$ and hence underestimated acuity. This would be the case if the amount of extraneous noise was approximately the same for all data sets. However, it is also possible that most data sets had low levels of extraneous noise, while a few data sets had very high levels. It is necessary to distinguish between these possibilities in order to estimate the bias of the group average.

To estimate the fraction of data sets that had high levels of extraneous noise, the difference between the maximum likelihood acuity estimate and the original acuity estimate was computed for each data set. The average differences were 0.1 ± 0.7 log unit for the constant stimulus infant data ($t = 0.71$, $p > .2$), 0.2 ± 0.4 octave for the staircase infant data ($t = 1.58$, $p > .1$), and 0.0 ± 0.2 octave for the staircase toddler data ($t = 0.42$, $p > .5$), none of which were statistically significant. This indicates that most of the acuity estimates could not have been significantly affected by extraneous noise. Although there was no overall tendency for the maximum likelihood acuity estimate to be different from the original acuity estimate, the individual data sets showed small but significant differences between the two acuity estimates. The absolute value of the difference between the two acuity estimates was 0.5 ± 0.5 octave for the constant stimulus infant data, 0.3 ± 0.3 octave for the staircase infant data, and 0.1 ± 0.1 octave for the staircase toddler data (these values were all statistically different from zero; $t > 14$, $p < .001$). The difference between acuity estimates was correlated with γ: For data sets with γ_{est} less than 100%, the original acuity estimate tended to be lower than the present estimate ($r = .56$ for the constant stimulus infant data, $r = .46$ for the staircase infant data, $r = .32$ for the staircase toddler data sets; in all cases, $p < .001$). This indicates that the populations were similar to the hypothetical population shown in the bottom right in Figure 6, with most data sets having little extraneous noise, but a few data sets having a high amount of extraneous noise.

Independence of Estimates of Extraneous and Stimulus-Related Noise

The analysis leading to Equation 3 assumes that extraneous noise (γ) and stimulus-related noise (β) are independent factors affecting the psychometric function. However, in fitting individual data sets, it is possible that γ and β may interact. For instance, if threshold is within an octave or two of the most intense stimulus used, and the fraction correct never reaches 100%, this could be due to low values of either β (shallow slope) or γ (upper asymptote below zero). This type of interaction would allow γ_{est} to be high when β_{est} is low, and vice versa. If such interactions occurred in the data analysis, they should be most obvious for the infant constant stimuli and staircase grating acuity data sets, which had the lowest mean values for γ_{est}. Figure 8 shows log (β_{est}) as a function of γ_{est} for these data sets; the correlations were $r = .30$ ($p < .001$) for staircase and $r = .15$ ($p < .02$) for constant stimuli. In both cases, there is a tendency for log β_{est} to be slightly smaller when γ_{est} is near 100%. However, γ_{est} accounts for little of the variance in log β_{est} (<10% staircase, <3% constant stimuli). Therefore, stimulus-related noise and extraneous noise can be considered relatively independent influences on performance.

Both stimulus-related noise and extraneous noise were relatively independent of age. For the infant acuity data sets, with both constant stimuli and staircase methods, there were no significant correlations with age for either β_{est} or γ_{est} ($r < .085$ in all cases, $p > .10$). For the operant data sets, β_{est} was slightly smaller at older ages ($r = .148$, $p < .01$), but this accounted for less than 3% of the variance. The operant data showed no dependence of γ_{est} on age ($r = .08$, $p > .10$) because most had values of $\gamma_{est} = 100\%$ and those with $\gamma_{est} < 100\%$ were distributed approximately evenly across the range of ages tested. There was not a significant difference in the mean values of β_{est} for the infant and toddler staircase grating acuity data sets, ($t = 0.50$, $p > .20$). For the infant contrast-sensitivity data sets, there was not a significant change with age for either β_{est} ($r = .06$, $p > .2$) or γ_{est} ($r = .12$, $p > .1$).

Consequences of Assuming That $\gamma = 100\%$

Infant grating acuity data sets tended to have maximum likelihoods with $\gamma_{est} < 100\%$. However, in many threshold estimation algorithms, it is assumed that the upper asymptote is always 100%. The consequences of this assumption were evaluated by obtaining maximum likelihood fits with γ_{est} fixed at 100% for the constant stimuli and staircase infant grating acuity data sets. For γ_{est} fixed at 100%, the highest likelihoods were usually obtained with shallower slopes (smaller β_{est}) and lower acuities (smaller α_{est}) than for γ_{est} allowed to vary. In addition, the maximum likelihood across (α_{est}, β_{est}) is usually greater when γ_{est} is allowed to vary than when γ_{est} is fixed

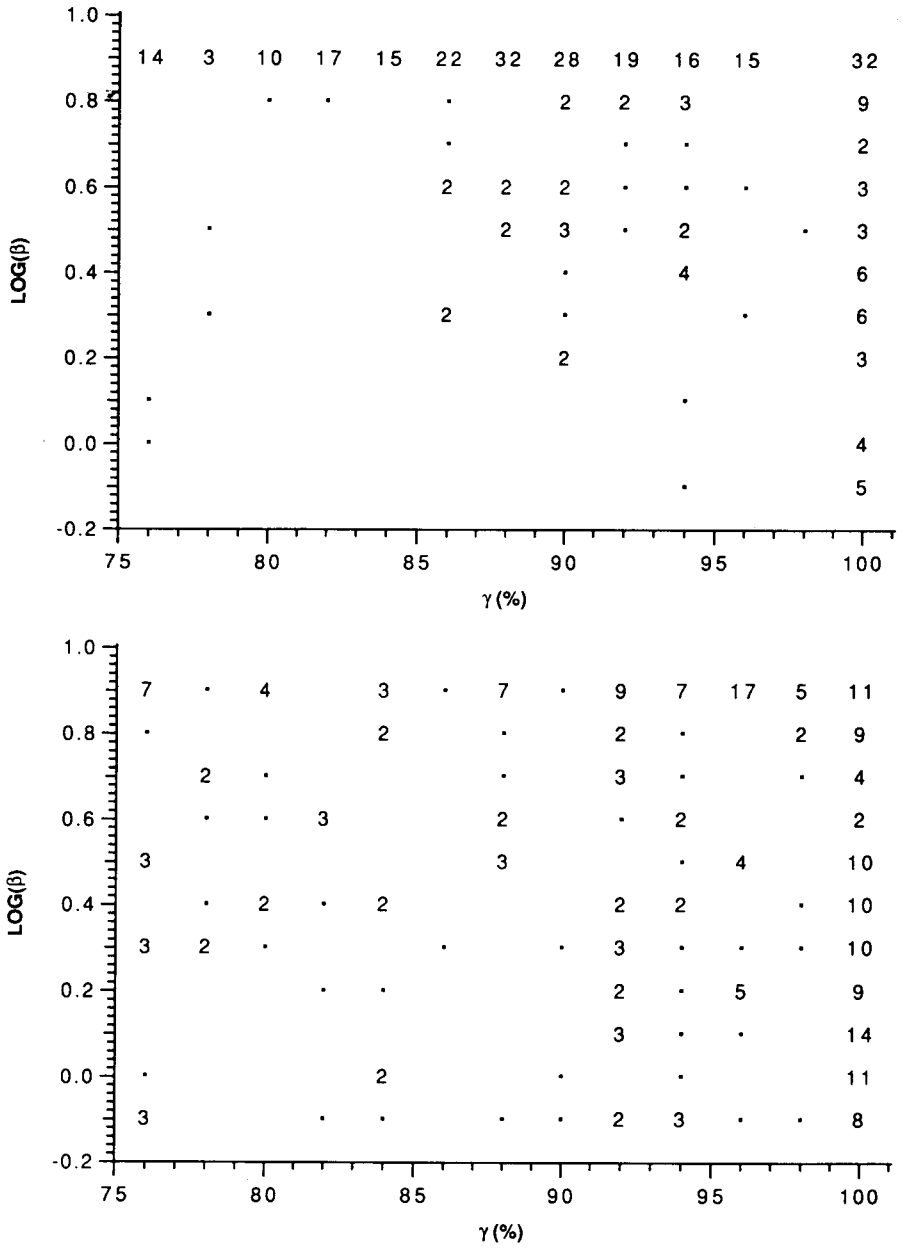

Figure 8. Scatterplots of extraneous noise (γ_{est}) versus slope of the psychometric function (log β_{est}) for infant acuity data gathered with staircase (upper panel) and constant stimuli (lower panel) methods. Because of the discrete parameter values used in the fitting procedure (10 for log β_{est}, 12 for log γ_{est}), many of the points in each graph represent more than one data set; in such cases, the number of data sets with each parameter set is indicated.

at 100%. For the combined constant stimuli and staircase data sets, fits obtained with γ_{est} fixed at 100% yielded lower acuities (0.6 octave constant stimuli, 1.4 octave staircase), lower likelihoods (0.4 log unit constant stimuli, 1.1 log unit staircase), and lower values of β_{est} (0.4 log unit constant stimuli, 0.7 log unit staircase) than did the fits that allowed γ_{est} to vary. These differences were all statistically significant ($t > 5$, $p < .001$).

Consequences of Assuming That Stimulus-Related Noise (β) Is Constant

The introduction of γ as a third variable in the psychometric function complicates data analysis. It has been suggested that this complexity could be reduced by fixing β (Klein & Manny, 1989). To evaluate the effects of this simplification, the 570 infant acuity data were reanalyzed with β fixed either at its average value (4.9) or at a lower

value (2.0), and the resulting threshold estimates were compared with those obtained when β was allowed to vary. For the constant stimuli data sets with β_{est} fixed at 4.9, the fraction of data sets with more than 1.0 octave difference in acuity was 33%, whereas with β_{est} fixed at 2.0, less than 10% of the data sets had differences of more than 0.5 octave. For the staircase data sets, less than 3% of the data sets had differences of more than 0.5 octave for either value of β_{est}. In most cases for $\beta_{est} = 2.0$, there was either no change in γ_{est} (69% for constant stimuli, 58% for staircase) or only a change in γ_{est} by 2% (17% constant stimuli, 23% staircase). Similarly, in most cases for $\beta_{est} = 2$, the likelihood decreased by either less than 0.5 log unit (91% constant stimuli, 59% staircase) or between 0.5 and 1.0 log unit (7% constant stimuli, 32% staircase). Overall, fixing $\beta_{est} = 2.0$ caused relatively little change in the fits.

Extraneous Noise in Test–Retest and Interocular Comparisons

Since analyses that do not account for extraneous noise (i.e., γ_{est} fixed at 100%) can yield acuity values significantly different from those obtained when extraneous noise is accounted for (i.e., γ_{est} is allowed to vary), the magnitudes of test-retest differences and interocular differences may be overestimated if changes in extraneous noise between tests are not accounted for. Infant staircase grating acuity data on monocular test-retest differences (38 pairs of tests) and on interocular differences (79 pairs) were examined. Test-retest differences were 0.4 octave higher when γ_{est} was fixed at 100% than when γ_{est} was allowed to vary (0.8 ± 0.7 octave vs. 0.4 ± 0.4 octave). Interocular differences were 0.3 octave higher when γ_{est} was fixed at 100% than when γ_{est} was allowed to vary (0.8 ± 0.7 octave vs. 0.5 ± 0.5 octave). In each case, these overestimates were statistically significant ($t > 2.9$, $p < .001$).

Comment

This analysis of published data from infants and toddlers indicates that extraneous noise can be significant and that the fraction of data sets affected by extraneous noise can be reduced by manipulating the experimental situation. Stimulus-related noise and extraneous noise are relatively independent factors. Fixing γ at 100% resulted in underestimates of acuity, whereas fixing β had little effect on data analysis. Analysis that does not consider extraneous noise may underestimate sensitivity, underestimate test-retest reliability, and overestimate interocular differences.

CONCLUSIONS

Simulations show that when extraneous noise is high, thresholds tend to be underestimated because of inappropriate stimulus distributions. Analysis of published data shows that extraneous noise can be a significant factor, increasing apparent test-retest and interocular differences. Simulations indicate that the most fruitful approach is to develop strategies that reduce or effectively eliminate extraneous noise, and analysis of published data shows that such strategies can be developed. It is difficult to estimate either the amount of extraneous noise affecting individual data sets or the slope of individual psychometric functions, but analysis of group data can yield useful criteria for rejecting individual data sets as unreliable *if* the overall estimate of extraneous noise is low for the group.

REFERENCES

BIRCH, E. E. (1985). Infant interocular acuity differences and binocular vision. *Vision Research*, **25**, 571-576.

BIRCH, E. E., & HALE, L. A. (1988). Criteria for monocular acuity deficit in infancy and early childhood. *Investigative Ophthalmology & Visual Science*, **29**, 636-643.

BROWN, A. M, DOBSON, V., & MAIER, J. (1987). Visual acuity of human infants at scotopic, mesopic and photopic luminances. *Vision Research*, **27**, 1845-1858.

DOBSON, V. (1983). Clinical applications of preferential looking measures of visual acuity. *Behavioral Brain Research*, **10**, 25-28.

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

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

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

KLEIN, S. A., & MANNY, R. E. (1989). Efficient estimation of thresholds with a small number of trials. *Noninvasive Assessment of the Visual System* (1989 Technical Digest Series), **7**, 80-83. Washington, DC: Optical Society of America.

LUSTED, L. B. (1967). Introduction to medical decision making. In B. Jacobson (Ed.), *Digest of the 7th International Conference on Medical and Biological Engineering* (p. 297). Stockholm: Almqvist & Wiksell.

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

MANNY, R. E., & KLEIN, S. A. (1985). A three alternative tracking paradigm to measure vernier acuity of older infants. *Vision Research*, **25**, 1245-1252.

MAYER, D. L., & DOBSON, V. (1982). Visual acuity development in infants and young children, as assessed by operant preferential looking. *Vision Research*, **22**, 1141-1151.

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.

METZ, C. E. (1982). ROC methodology in radiologic imaging. *Investigative Radiology*, **21**, 720-733.

NACHMIAS, J. (1982). Starting point bias of a recent psychophysical method. *American Journal of Optometry & Physiological Optics*, **59**, 845-847.

O'REGAN, J. K., & HUMBERT, R. (1989). Estimating psychometric functions in forced-choice situations: Significant biases found in threshold and slope estimations when small samples are used. *Perception & Psychophysics*, **46**, 434-442.

PELLI, D., ROBSON, J. G., & WILKINS, A. J. (1988). The design of a new letter contrast chart for measuring contrast sensitivity. *Clinical Vision Sciences*, **2**, 187-199.

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

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

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

SWANSON, W. H., & BIRCH, E. E. (1990). Infant spatiotemporal vision: Dependence of spatial contrast sensitivity on temporal frequency. *Vision Research*, **30**, 1033-1048.

TAYLOR, M. M., & CREELMAN, C. D. (1967). PEST: Efficient estimates on probability functions. *Journal of the Acoustical Society of America*, **41**, 782-787.

TELLER, D. Y., MAR, C., & PRESTON, K. L. (in press). Statistical properties of 500-trial infant psychometric functions. In L. Werner & E. Rubel (Eds.), *Developmental psychoacoustics*. Washington, DC: American Psychological Association.

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

WATSON, A. B., & FITZHUGH, A. (1990). The method of constant stimuli is inefficient. *Perception & Psychophysics*, **47**, 87-91.

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

NOTES

1. For 2AFC, the subset of trials affected by extraneous noise will be $2(1-\gamma)$ of the total number of trials. On these trials, the response will be correct 50% of the time by chance, yielding $0.5[2(1-\gamma)]$. The right side of Equation 1 will only apply to the subset of trials not affected by extraneous noise, which will be $1-2(1-\gamma)$, yielding $\{P(x)+0.5[1-P(x)]\}\{1-2(1-\gamma)\}$. Adding these fractions correct for the two subsets of trials yields

$$R(x) = 0.5[2(1-\gamma)] + \{P(x)+0.5[1-P(x)]\}\,[1-2(1-\gamma)].$$

Algebraic rearrangement yields Equation 3.

2. This method estimates the 71% correct point of $P(x)$, so the mean of reversals tends to be slightly lower than α, which is the 75% correct point. For the analysis of published data, the original thresholds were compared with the 71% correct point for the best-fitting $P(x)$.

3. The staircases were terminated when a required number of reversals were obtained; the average number of trials required was about 20 when threshold was near the starting point and about 45 when threshold was 3 log units below the starting point.

4. The reason that the bias for γ_{est} was diminished more by the second strategy than by the first is that when all three parameters are allowed to vary there can be interaction between β_{est} and γ_{est}; that is, shallow slopes may allow higher upper asymptotes. This interaction is discussed further in the section titled "Independence of estimates of extraneous and stimulus-related noise."

5. ROC methodology assumes that there are two Gaussian distributions to be distinguished. For each strategy and population, the z scores for false positives were plotted as a function of the z scores for the true positives, for all of the criteria tested. These scatterplots were well described by straight lines, with $r > .98$ for all cases except one (Strategy 5 with Population 3 had $r = .87$; removing the highest and lowest points from the analysis yielded $r = .99$), indicating that the assumption of two Gaussian distributions was appropriate. The slope and intercepts of the best-fitting lines were used to generate the smooth ROC curves shown in the upper portion of Figure 6. Since the slopes were always greater than 1.0 and varied from one scatterplot to the next, d' is not an appropriate measure of success; therefore, the area under the smooth ROC curves was used.

(Manuscript received April 8, 1991;
revision accepted for publication November 26, 1991.)