Perception & Psychophysics
2001, 63 (8), 1421-1455

Measuring, estimating, and understanding the psychometric function: A commentary

STANLEY A. KLEIN
University of California, Berkeley, California

The psychometric function, relating the subject's response to the physical stimulus, is fundamental to psychophysics. This paper examines various psychometric function topics, many inspired by this special symposium issue of *Perception & Psychophysics*: What are the relative merits of objective yes/no versus forced choice tasks (including threshold variance)? What are the relative merits of adaptive versus constant stimuli methods? What are the relative merits of likelihood versus up–down staircase adaptive methods? Is 2AFC free of substantial bias? Is there no efficient adaptive method for objective yes/no tasks? Should adaptive methods aim for 90% correct? Can adding more responses to forced choice and objective yes/no tasks reduce the threshold variance? What is the best way to deal with lapses? How is the Weibull function intimately related to the d' function? What causes bias in the likelihood goodness-of-fit? What causes bias in slope estimates from adaptive methods? How good are nonparametric methods for estimating psychometric function parameters? Of what value is the psychometric function slope? How are various psychometric functions related to each other? The resolution of many of these issues is surprising.

Psychophysics, as its metaphysical sounding name indicates, is the scientific discipline that explores the connection between physical stimuli and subjective responses. The psychometric function (PF) provides the fundamental data for psychophysics, with the PF abscissa being the stimulus strength and the ordinate measuring the observer's response. I shudder when I think about the many hours researchers (including myself) have wasted in using inefficient procedures to measure the PF, as well as when I see procedures being used that do not reveal all that could be extracted with the same expenditure of time—and, of course, when I see erroneous conclusions being drawn because of biased methodologies. The articles in this special symposium issue of *Perception & Psychophysics* deal with these and many more issues concerning the PF. I was a reviewer on all but one of these articles and have watched them mature. I have now been given the opportunity to comment on them once more. This time I need not quibble with minor items but rather can comment on several deeper issues. My commentary is divided into three sections.

1. What is the PF and how is it specified? Inspired by Strasburger's article (Strasburger, 2001a) on a new definition of slope as applied to a wide variety of PF shapes, I will comment on several items: the connections between several forms of the psychometric function (Weibull, cumulative normal, and d'); the relationship between slope of the PF with a linear versus a logarithmic abscissa; and the connection between PFs and signal detection theory.

2. What are the best experimental techniques for *measuring* the PF? In most of the articles in this symposium, the two-alternative forced choice (2AFC) technique is used to measure threshold. The emphasis on 2AFC is appropriate; 2AFC seems to be the most common methodology for this purpose. Yet for the same reason I shall raise questions about the 2AFC technique. I shall argue that both its limited ability to reveal underlying processes and its inefficiency should demote it from being the method of choice. Kaernbach (2001b) introduces an "unforced choice" method that offers an improvement over standard 2AFC. The article by Linschoten, Harvey, Eller, and Jafek (2001) on measuring thresholds for taste and smell is relevant here: Because each of their trials takes a long time, an optimal methodology is needed.

3. What are the best analytic techniques for *estimating* the properties of PFs once the data have been collected? Many of the papers in this special issue are relevant to this question.

The articles in this special issue can be divided into two broad groups: Those that did not surprise me, and those that did. Among the first group, Leek (2001) gives a fine historical overview of adaptive procedures. She also provides a fairly complete list of references in this field. For details on modern statistical methods for analyzing data, the pair of papers by Wichmann and Hill (2001a, 2001b) offer an excellent tutorial. Linschoten et al. (2001) also provide a good methodological overview, comparing different methods for obtaining data. However, since these articles were

I thank all the authors contributing to this special issue for their fine papers. I am especially grateful for the many wonderful suggestions and comments that I have had from David Foster, Lew Harvey, Christian Kaernbach, Marjorie Leek, Neil Macmillan, Jeff Miller, Hans Strasburger, Bernhard Treutwein, Christopher Tyler, and Felix Wichmann. This research was supported by Grant R01EY04776 from the National Institutes of Health. Correspondence should be addressed to S. A. Klein, School of Optometry, University of California, Berkeley, CA 94720-2020.

nonsurprising and for the most part did not shake my previous view of the world, I feel comfortable with them and feel no urge to shower them with words. On the other hand, there were a number of articles in this issue whose results surprised me. They caused me to stop and consider whether my previous thinking had been wrong or whether the article was wrong. Among the present articles, six contained surprises: (1) the Miller and Ulrich (2001) nonparametric method for analyzing PFs, (2) the Strasburger (2001b) finding of extremely steep psychometric functions for letter discrimination, (3) the Strasburger (2001a) new definition of psychometric function slope, (4) the Wichmann and Hill (2001a) analysis of biased goodness-of-fit and bias due to lapses, (5) the Kaernbach (2001a) modification to 2AFC, and (6) the Kaernbach (2001b) analysis of why staircase methods produce slope estimates that are too steep. The issues raised in these articles are instructive, and they constitute the focus of my commentary. Item 3 is covered in Section I, Item 5 will be discussed in Section II, and the remainder are discussed in Section III. A detailed overview of the main conclusions will be presented in the summary at the end of this paper. That might be a good place to begin.

When I began working on this article, more and more threads captured my attention, causing the article to become uncomfortably large and diffuse. In discussing this situation with the editor, I decided to split my article into two publications. The first of them is the present article, focused on the nine articles of this special issue and including a number of overview items useful for comparing the different types of PFs that the present authors use. The second paper will be submitted for publication in *Perception & Psychophysics* in the future (Klein, 2002).

The articles in this special issue of *Perception & Psychophysics* do not cover all facets of the PF. Here I should like to single out four earlier articles for special mention: King-Smith, Grigsby, Vingrys, Benes, and Supowit (1994) provide one of the most thoughtful approaches to likelihood methods, with important new insights. Treutwein's (1995) comprehensive, well-organized overview of adaptive methods should be required reading for anyone interested in the PF. Kontsevich and Tyler's (1999) adaptive method for estimating both threshold and slope is probably the best algorithm available for that task and should be looked at carefully. Finally, the paper with which I am most familiar in this general area is McKee, Klein, and Teller's (1985) investigation of threshold confidence limits in probit fits to 2AFC data. In looking over these papers, I have been struck by how Treutwein (1995), King-Smith et al. (1994), and McKee et al. (1985) all point out problems with the 2AFC methodology, a theme I will continue to address in Section II of this commentary.

I. TYPES OF PSYCHOMETRIC FUNCTIONS

The Probability-Based (High-Threshold) Correction for Guessing

The PF is commonly written as follows:

$$P(x) = \gamma + (1-\lambda-\gamma)p(x), \tag{1A}$$

where $\gamma = P(0)$ is the lower asymptote, $1-\lambda$ is the upper asymptote, $p(x)$ is the PF that goes from 0% to 100%, and $P(x)$ is the PF representing the data that goes from γ to $1-\lambda$. The stimulus strength, x, typically goes from 0 to a large value for detection and from large negative to large positive values for discrimination. For discrimination tasks where $P(x)$ can go from 0% (for negative stimulus values) to 100% (for positive values), there is, typically, symmetry between the negative and positive range so that $\lambda = \gamma$. Unless otherwise stated, I will, for simplicity, ignore lapses (errors made to perceptible stimuli) and take $\lambda = 0$ so that $P(x)$ becomes

$$P(x) = \gamma + (1-\gamma)p(x). \tag{1B}$$

In the Section III commentary on Strasburger (2001b) and Wichmann and Hill (2001a), I will discuss the benefit of setting the lapse rate, λ, to a small value (like $\lambda = 1\%$) rather than 0% (or the 0.01% value that Strasburger used) to minimize slope bias.

When coupled with a high threshold assumption, Equation 1B is powerful in connecting different methodologies. The high threshold assumption is that the observer is in one of two states: detect or not detect. The detect state occurs with probability $p(x)$. If the stimulus is not detected then one guesses with the guess rate, γ. Given this assumption, the percent correct will be

$$P(x) = p(x) + \gamma\left[1-p(x)\right], \tag{1C}$$

which is identical to Equation 1B. The first term corresponds to the occasions when one detects the stimulus, and the second term corresponds to the occasions when one does not.

Equation 1 is often called the *correction for guessing* transformation. The correction for guessing is clearer if Equation 1B is rewritten as:

$$p(x) = \frac{P(x) - P(0)}{1 - P(0)}. \tag{2}$$

The beauty of Equation 2 together with the high threshold assumption is that even though $\gamma = P(0)$ can change, the fundamental PF, $p(x)$, is unchanged. That is, one can alter γ by changing the number of alternatives in a forced choice task [γ = 1/(number of alternatives)], or one can alter the false alarm rate in a yes/no task; $p(x)$ remains unchanged and recoverable through Equation 2.

Unfortunately, when one looks at actual data, one will discover that $p(x)$ does change as γ changes for both yes/no and forced choice tasks. For this and other reasons, the high threshold assumption has been discredited. A modern method, signal detection theory, for doing the correction for guessing is to do the correction after a z-score transformation. I was surprised that signal detection theory was barely mentioned in any of the articles constituting this special issue. I consider this to be sufficiently important that I want to clarify it at the outset. Before I can introduce the newer approach to the correction for guessing, the connection between probability and z-score is needed.

The Connection Between Probability and *z*-Score

The Gaussian distribution and its integral, the cumulative normal function, play a fundamental role in many approaches to PFs. The cumulative normal function, $\Phi(z)$, is the function that connects the *z*-score (z) to probability (prob):

$$\text{prob} = \Phi(z) = (2\pi)^{-0.5} \int_{-\infty}^{z} dy \exp\left(-\frac{y^2}{2}\right). \tag{3A}$$

The function $\Phi(z)$ and its inverse are available in programs such as Excel, but not in Matlab. For Matlab, one must use:

$$\text{prob} = \Phi(z) = \frac{1+\text{erf}\left(-\frac{z}{\sqrt{2}}\right)}{2}. \tag{3B}$$

where the error function,

$$\text{erf}(x) = 2\pi^{-0.5} \int_{0}^{x} dy \exp\left(-y^2\right),$$

is a standard Matlab function. I usually check that $\Phi(-1) = 0.1587$, $\Phi(0) = 0.5$, and $\Phi(1) = 0.8413$ in order to be sure that I am using erf properly. The inverse cumulative normal function, used to go from prob to z is given by

$$z = \Phi^{-1}(\text{prob}) = \sqrt{2}\text{erfinv}(2 * \text{prob} - 1). \tag{4}$$

Equations 3 and 4 do not specify whether one uses prob $= p$ or prob $= P$ (see Equation 1 for the distinction between the two). In this commentary, both definitions will be used, with the choice depending on how one does the correction for guessing—our next topic. The choice prob $= p(x)$, means that a cumulative normal function is being used for the underlying PF and that the correction for guessing is done as in Equations 1 and 2. On the other hand, in a yes/no task, if we choose prob $= P(x)$, then one is doing the *z*-score transform of the PF data before the correction for guessing. As is clarified in the next section, this procedure is a signal detection "correction for guessing" and the PF will be called the d' function. The distinction between the two methods for correction for guessing has generated confusion and is partly responsible for why many researchers do not appreciate the simple connection between the PF and d'.

The *z*-Score Correction for Bias and Signal Detection Theory: Yes/No

Equation 2 is a common approach to correction for guessing in both yes/no and forced choice tasks, and it will be found in many of the articles in this special issue. In a yes/no method for detection, the lower asymptote, $P(0) = \gamma$, is the false alarm rate, the probability of saying "yes" when a blank stimulus is present. In the past, one instructed subjects to keep γ low. A few blank catch trials were included to encourage subjects to maintain their low false alarm rate. Today, the correction is done using a *z*-score ordinate and it is now called the *correction for response bias*, or simply the *bias correction*. One uses as many trials at the zero level as at other levels, one encourages more false alarms, and the framework is called *signal detection theory*.

The *z*-score (d') correction for bias provides a direct, but not well appreciated, connection between the yes/no psychometric function and the signal detection d'. The two steps are as follows: (1) Convert the percent correct, $P(x)$, to z scores using Equation 4. (2) Do the correction for bias by choosing the zero point of the ordinate to be the z score for the point $x = 0$. Finally, give the ordinate the name, d'. This procedure can be written as:

$$d'(x) = z(x) - z(0). \tag{5}$$

For a detection task, $z(x)$ is called the z score of the hit rate and $z(0)$ is called the z score of the lower asymptote (γ), or the false alarm rate. It is the false alarm rate because the stimulus at $x = 0$ is the blank stimulus. In order to be able to do this correction for bias accurately, one must put as many trials at $x = 0$ as one puts at the other levels. Note the similarity of Equations 2 and 5. The main difference between Equations 2 and 5 is whether one makes the correction in probability or in z score. In order to distinguish this approach from the older yes/no approach associated with high threshold theory, it is often called the *objective yes/no method*, where "objective" means that the response bias correction of Equation 5 is used.

Figure 1 and its associated Matlab Code 1 in the Appendix illustrates the process represented in Equation 2. Panel a is a Weibull PF on a linear abscissa, to be introduced in Equation 12. Panel b is the z score of panel a. The lower asymptote is at $z = -1$, corresponding to $P = 15.87\%$. For now, the only important point is that in panel b, if one measures the curve from the bottom of the plot ($z = -1$), then the ordinate becomes d' because $d' = z - z(0) = z + 1$. Panels d and e are the same as panels a and b, except that instead of a linear abscissa they have natural log abscissas. More will be said about these figures later.

I am ignoring for now the interesting question of what happens to the shape of the psychometric function as one changes the false alarm rate, γ. If one uses multiple ratings rather than the binary yes/no response, one ends up with $M - 1$ PFs for M rating categories, and each PF has a different γ. For simplicity, this paper assumes a unity ROC slope, which guarantees that the d'function is independent of γ. The ROC slopes can be measured using an objective yes/ no method as mentioned in Section II in the list of advantages of the yes/no method over the forced choice method.

I bring up the d'function (Equation 5) and signal detection theory at the very beginning of this commentary because it is an excellent methodology for measuring thresholds efficiently; it can easily be extended to the suprathreshold regime (it does not saturate at $P = 1$), and it has a solid theoretical underpinning. Yet it is barely mentioned in any of the articles in this issue. So the reader needs to keep in mind that there is an alternative approach to PFs. I would strongly recommend the book *Detection Theory: A User's Guide* (Macmillan & Creel-

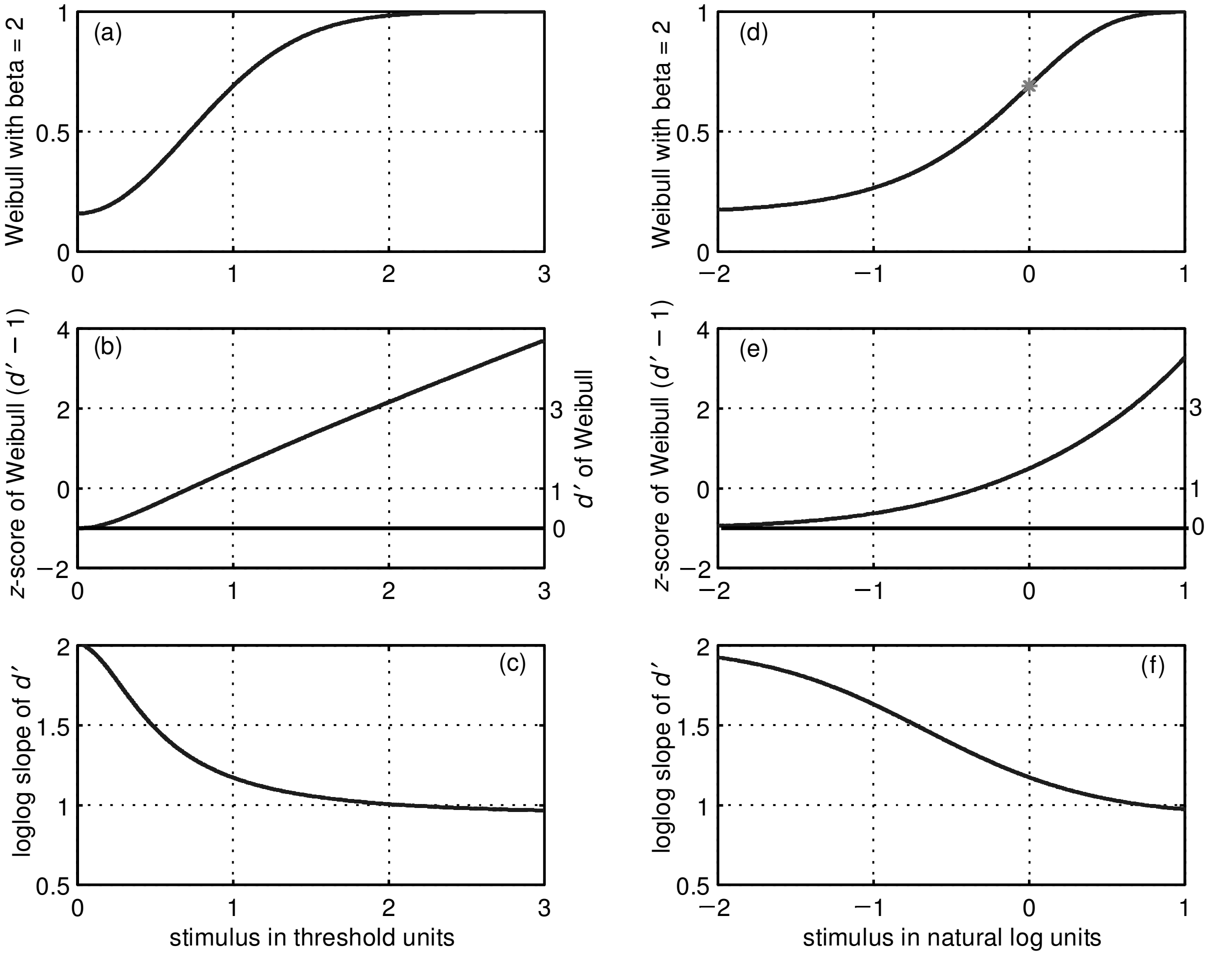

Figure 1. Six views of the Weibull function: $P_{\text{weibull}} = 1 - (1-\gamma)\exp(-x_t^{\beta})$, where $\gamma = 0.1587$, $\beta = 2$, and x_t is the stimulus strength in threshold units. Panels a–c have a linear abscissa, with $x_t = 1$ being the threshold. Panels d–f have a natural log abscissa, with $y_t = 0$ being the threshold. In panels a and d, the ordinate is probability. The asterisk in panel d at $y_t = 0$ is the point of maximum slope on a logarithmic abscissa. In panels b and e, the ordinate is the z score of panels a and d. The lower asymptote is $z = -1$. If the ordinate is redefined so that the origin is at the lower asymptote, the new ordinate, shown on the right of panels b and e, is $d'(x_t) = z(x_t) - z(0)$, corresponding to the signal detection d' for an objective yes/no task. In panels c and f, the ordinate is the log–log slope of d'. At $x_t = 0$, the log–log slope = β. The log–log slope falls rapidly as the stimulus strength approaches threshold. The Matlab program that generated this figure is Appendix Code 1.

man, 1991) for anyone interested in the signal detection approach to psychophysical experiments (including the effect of nonunity ROC slopes).

The z-Score Correction for Bias and Signal Detection Theory: 2AFC

One might have thought that for 2AFC the connection between the PF and d' is well established—namely (Green & Swets, 1966),

$$d'(x) = z(x)\sqrt{2}\,, \tag{6}$$

where $z(x)$ is the z score of the average of $P_1(x)$ for correct judgments in Interval 1 and $P_2(x)$ for correct judgments in Interval 2. Typically this average $P(x)$ is calculated by dividing the total number of correct trials by the total number of trials. It is generally assumed that the 2AFC procedure eliminates the effect of response bias on threshold. However, in this section I will argue that the d' as defined in Equation 6 is affected by an interval bias, when one interval is selected more than the other.

I have been thinking a lot about response bias in 2AFC tasks because of my recent experience as a subject in a temporal 2AFC contrast discrimination study, where contrast is defined as the change in luminance divided by the background luminance. In these experiments, I noticed that I had a strong tendency to choose the second interval more than the first. The second interval typically appears subjectively to be about 5% higher in contrast than it really is. Whether this is a perceptual effect because the intervals are too close together in time (800 msec) or a cognitive effect does not matter for the present article. What does matter is that this bias produces a downward bias in d'. With feedback, lots of practice, and lots of experience being a subject, I was able to reduce this interval bias and equalize the number of times I responded with each interval. Naïve subjects may have a more difficult time reducing the bias. In this section, I show that there is a very simple method for removing the interval bias, by converting the 2AFC data to a PF that goes from 0% to 100%.

The recognition of bias in 2AFC is not new. Green and Swets (1966), in their Appendix III.3.4, point out that the bias in choice of interval does result in a downward bias in d'. However, they imply that the effect of this bias is small and can typically be ignored. I should like to question that implication, by using one of their own examples to show that the bias can be substantial.

In a run of 200 trials, the Green and Swets example (Green & Swets, 1966, p. 410) has 95 out of 100 correct when the test stimulus is in the second interval (z_2 = 1.645) and 50 out of 100 correct (z_1 = 0) when the stimulus is in the first interval. The standard 2AFC way to analyze these data would be to average the probabilities (95% + 50%)/ 2 = 72.5% correct (z_{correct} = 0.598), corresponding to $d' = z\sqrt{2} = 0.845$. However, Green and Swets (p. 410) point out that according to signal detection theory one should analyze this data by averaging the z scores rather than averaging the probabilities, or

$$d' = \sqrt{2}\,\frac{\left(z_2 + z_1\right)}{2} = \frac{1.645}{\sqrt{2}} = 1.163\,. \tag{7}$$

The ratio between these two ways of calculating d' is 1.163/0.845 = 1.376. Since d' is approximately linearly related to signal strength in discrimination tasks, this 38% reduction in d' corresponds to an erroneous 38% increase in predicted contrast discrimination threshold, when one calculates threshold the standard way. Note that if there had been no bias, so that the responses would be approximately equally divided across the two intervals, then $z_2 \approx z_1$ and Equation 7 would be identical to the more familiar Equation 6. Since bias is fairly common, especially among new observers, the use of Equation 7 to calculate d' seems much more reasonable than using Equation 6. It is surprising that the bias correction in Equation 7 is rarely used.

Green and Swets (1966) present a different analysis. Instead of comparing d' values for the biased versus nonbiased conditions, they convert the d's back to percent correct. The corrected percent correct (corresponding to d' = 1.163) is 79.5%. In terms of percent correct, the bias seems to be a small effect, shifting percent correct a mere 7% from 72.5% to 79.5%. However, the d' ratio of 1.38 is a better measure of the bias since it is directly related to the error in discrimination threshold estimates.

A further comment on the magnitude of the bias may be useful. The preceding subsection discussed the criterion bias of yes/no methods, which contributes *linearly* to d' (Equation 5). The present section discusses the 2AFC interval bias that contributes *quadratically* to d'. Thus, for small amounts of bias, the decrease in d' is negligible. However, as I have pointed out, the bias can be large enough to make a significant contribution to d'.

It is instructive to view the bias correction for the full PF corresponding to this example. Cumulative normal PFs are shown in the upper left panel of Figure 2. The curves labeled C1 and C2 are the probability correct for the first and second intervals, I1 and I2. The asterisks correspond to the stimulus strength used in this example, at 50% and 95% correct. The dot–dashed line is the average of the two PFs for the individual intervals. The slope of the PF is set by the dot–dashed line (the averaged data) being at 50% (the lower asymptote) for zero stimulus strength. The lower left panel is the z-score version of the upper panel. The dashed line is the average of the two solid lines for z scores in I1 and I2. This is the signal detection method of averaging that Green and Swets (1966) present as the proper way to do the averaging. The dot–dashed line shown in the upper left panel is the z score of the average probability. Notice that the z score for the averaged probability is lower than the averaged z score, indicating a downward bias in d' due to the interval bias, as discussed at the beginning of this section. The right pair of panels are the same as the left pair except that instead of plotting C1, we plot 1−C1, the probability of responding I2 incorrectly. The dashed line is half the difference between the two solid lines. The other difference is that in the lower right panel we have multiplied the dashed and dash–dotted lines by $\sqrt{2}$ so that these lines are d' values rather than z scores.

The final step in dealing with the 2AFC bias is to flip the 1−C1 curve horizontally to negative abscissa values as in Figure 3. The ordinate is still the probability correct in interval 2. The abscissa becomes the difference in stimulus strengths between I2 and I1. The flipped branch is the probability of responding I2 when the I2 stimulus strength is less than that of I1 (an incorrect response). Figure 3a with the ordinate going from 0 to 100% is the proper way to represent 2AFC discrimination data. The other item that I have changed is the scale on the abscissa to show what might happen in a real experiment. The ordinate values of 50% and 95% for the Green and Swets (1966) example have been placed at a contrast difference of 5%. The negative 5% value corresponds to the case in which the positive test pattern is in the first interval. Threshold corresponds to the inverse of the PF slope. The bottom panel shows the standard signal detection representation of the signal in I1 and I2. d' is the distance between these symmetric stimuli in standard deviation units. The Gaussians are centered at ±5%. The vertical line at −5% is the criterion, such that 50% and 95% of the area of the two Gaussians is above the criterion. The z-score difference of 1.645 between the two Gaussians must be divided by $\sqrt{2}$ to get d', because each trial had two stimulus presentations with independent informa-

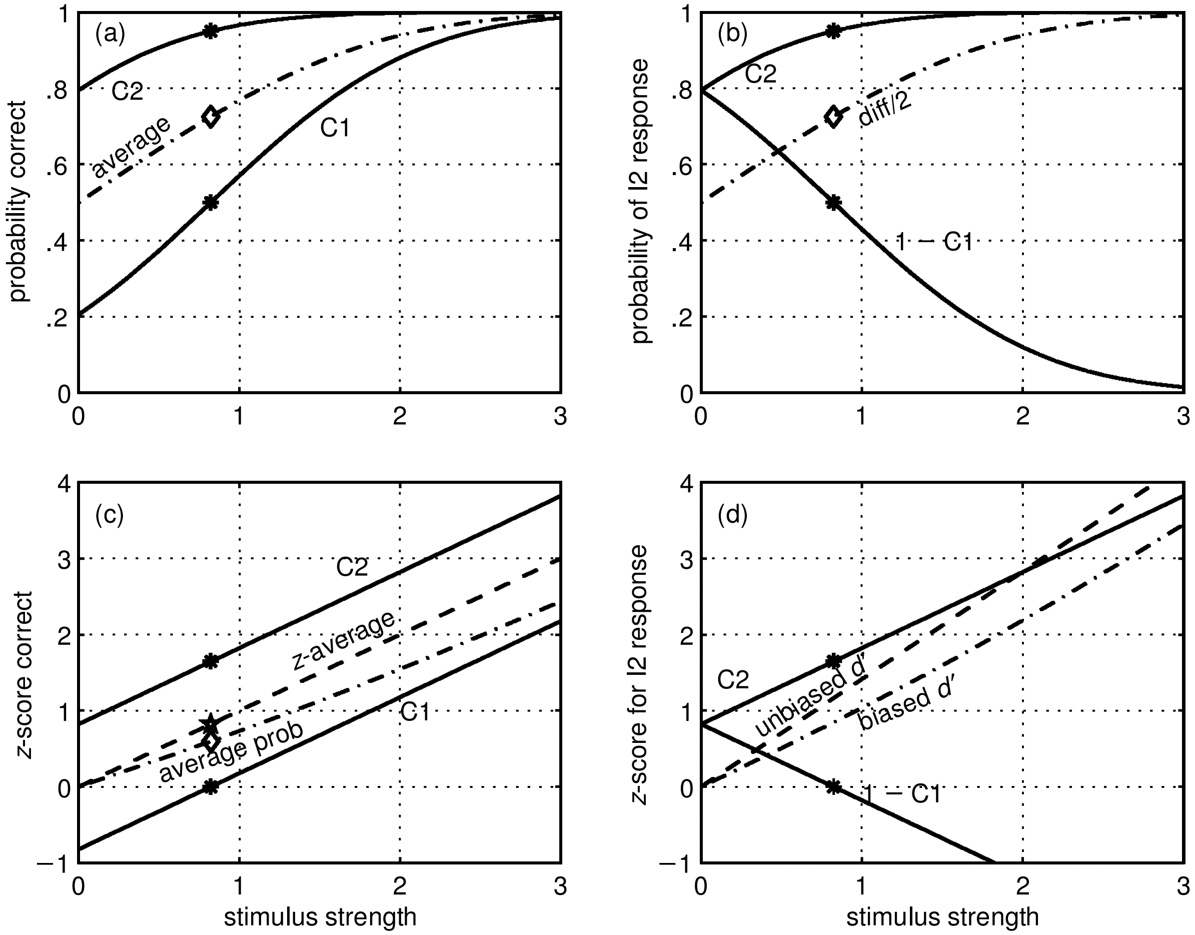

Figure 2. 2AFC psychometric functions with a strong bias in favor of responding Interval 2 (I2). The bias is chosen from a specific example presented by Green and Swets (1966), such that the observer has 50% and 95% correct when the test is in I1 and I2, respectively. These points are marked by asterisks. The psychometric function being plotted is a cumulative normal. In all panels, the abscissa is x_t, the stimulus strength. (a) The psychometric functions for probability correct in I1 and I2 are shown and labeled C1 and C2. The average of the two probabilities, labeled *average*, is the dot–dashed line; it is the curve that is usually reported. The diamond is at 72.5% the average percent correct of the two asterisks. (b) Same as panel a, except that instead of showing C1, we show 1–C1, the probability of saying I2 when the test was in I1. The abscissa is now labeled "probability of I2 response." (c) *z* scores of the three probabilities in panel a. An additional dashed line is shown that is the average of the C1 and C2 *z*-score curves. The diamond is the *z*-score of the diamond in panel a, and the star is the *z*-score average of the two panel c asterisks. (d) The sign of the C2 curve in panel c is flipped, to correspond to panel b. The dashed and dot–dashed lines of panel c have been multiplied by $\sqrt{2}$ in panel d so that they become *d′*.

tion for the judgment. This procedure, identical to Equation 7, gives the same *d′* as before.

Three Distinctions for Clarifying PFs

In dealing with PFs, three distinctions need to be made: yes/no versus forced choice, detection versus discrimination, and constant stimuli versus adaptive methods. These distinctions are usually clear, but I should like to point out some subtleties.

For the forced choice versus yes/no distinction, there are two sorts of forced choice tasks. The standard version has multiple intervals, separated spatially or temporally, and the stimulus is in only one of the intervals. In the other version, one of *N* stimuli is shown and the observer responds with a number from 1 to *N*. For example, Strasburger (2001b) presented 1 of 10 letters to the observer in a 10AFC task. Yes/no tasks have some similarity to the latter type of forced choice task. Consider, for example, a detection experiment in which one of five contrasts (including a blank) are presented to the observer and the observer responds with numbers from 1 to 5. This would be classified as a rating scale, method of constant stimuli, yes/no task, since only a single stimulus is presented and the rating is based on a one-dimensional intensity.

The detection/discrimination distinction is usually based on whether the reference stimulus is a natural zero point. For example, suppose the task is to detect a high spatial frequency test pattern added to a spatially identical reference pattern. If the reference pattern has zero contrast, the task is detection. If the reference pattern has a high contrast, the task is discrimination. Klein (1985) discusses these tasks in terms of monopolar and bipolar cues. For discrimination, a bipolar cue must be available whereby the test pattern can be either positive or negative in relation to the reference. If one cannot discriminate the negative cue from the positive, then it can be called a detection task.

Finally, the constant stimuli versus adaptive method distinction is based on the former's having preassigned test levels and the latter's having levels that shift to a desired placement. The output of the constant stimulus method is a full PF and is thus fully entitled to be included in this special issue. The output of adaptive methods is typically only a single number, the threshold, specifying the location, but not the shape, of the PF. Two of the papers in this issue (Kaernbach, 2001b; Strasburger, 2001b) explore the possibility of also extracting the slope from adaptive data that concentrates trials around one level. Even though adaptive methods do not measure much about the PF, they are so popular that they are well represented in this special issue.

The 10 rows of Table 1 present the articles in this special issue (including the present article). Columns 2–5 correspond to the four categories associated with the first two

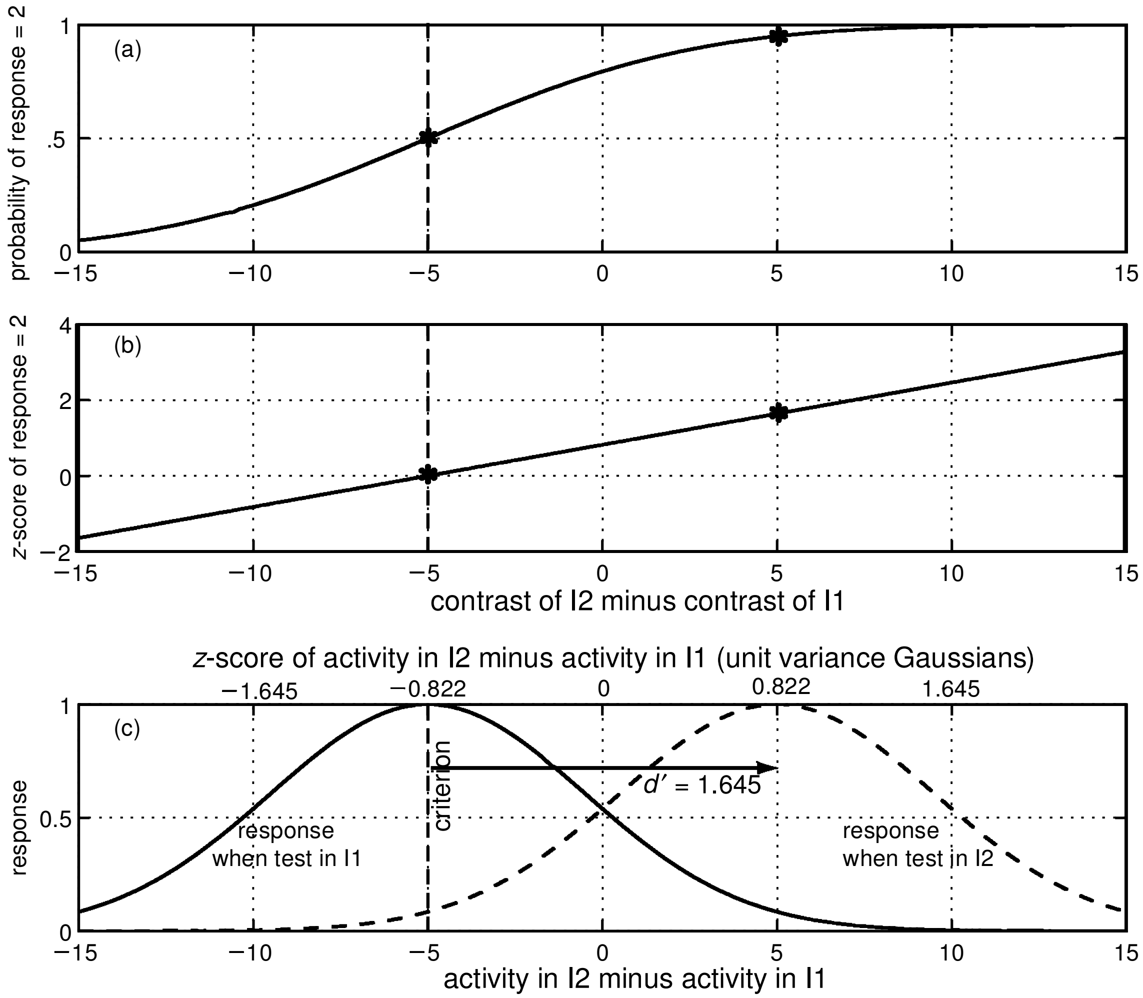

Figure 3. 2AFC discrimination PF from Figure 2 has been extended to the 0% to 100% range without rescaling the ordinate. In panels a and b, the two right-hand panels of Figure 2 have been modified by flipping the sign of the abscissa of the negative slope branch where the test is in Interval 1 (I1). The new abscissa is now the stimulus strength in I2 minus the strength in I1. The abscissa scaling has been modified to be in stimulus units. In this example, the test stimulus has a contrast of 5%. The ordinate in panel a is the probability that the response is I2. The ordinate in panel b is the *z* score of the panel a ordinate. The asterisks are at the same points as in Figure 2. Panel c is the standard signal detection picture when noise is added to the signal. The abscissa has been modified to being the activity in I2 minus the activity in I1. The units of activation have arbitrarily been chosen to be the same as the units of panels a and b. Activity distributions are shown for stimuli of −5 and +5 units, corresponding to the asterisks of panels a and b. The subject's criterion is at −5 units of activation. The upper abscissa is in *z*-score units, where the Gaussians have unit variance. The area under the two distributions above the criterion are 50% and 95% in agreement with the probabilities shown in panel a.

distinctions. The last column classifies the articles according to the adaptive versus constant stimuli distinction. In order to open up the full variety of PFs for discussion and to enable a deeper understanding of the relationship among different types of PFs, I will now clarify the interrelationship of the various categories of PFs and their connection to the signal detection approach.

Lack of adaptive method for yes/no tasks with controlled false alarm rate. In Section II, I will bring up a number of advantages of the yes/no task in comparison with to the 2AFC task (see also Kaernbach, 1990). Given those yes/no advantages, one may wonder why the 2AFC method is so popular. The usual answer is that 2AFC has no response bias. However, as has been discussed, the yes/no method allows an unbiased d' to be calculated, and the 2AFC method does allow an interval bias that affects d'. Another reason for the prevalence of 2AFC experiments is that a multitude of adaptive methods are available for 2AFC but barely any available for an objective yes/no task in which the false alarm rate is measured so that d' can be calculated. In Table 1, with one exception, the rows with adaptive methods are associated with the forced choice method. The exception is Leek (2001), who discusses adaptive yes/no methods in which no blank trials are presented. In that case, the false alarm rate is not measured, so d' cannot be calculated. This type of yes/no method does not belong in the "objective" category of concern for the present paper. The "1990" entries in Table 1 refer to Kaernbach's (1990) description of a staircase method for an objective yes/no task in which an equal number of blanks are intermixed with the signal trials. Since Kaernbach could have used that method for Kaernbach (2001b), I placed the "1990" entry in his slot.

Kaernbach's (1990) yes/no staircase rules are simple. The signal level changes according to a rule such as the following: Move down one level for correct responses: a hit, <yes|signal> or a correct rejection <no|blank>; move up three levels for wrong responses: a miss <no|signal> or a false alarm <yes|blank>. This rule, similar to the one down–three up rule used in 2AFC, places the trials so that the average of the hit rate and correct rejection rate is 75%. What is now needed is a mechanism to get the observer to establish an optimal criterion that equalizes the number of "yes" and "no" responses (the ROC negative diagonal). This situation is identical to the problem of getting 2AFC observers to equalize the number of responses to Intervals 1 and 2. The quadratic bias in d' is the same in both cases. The simplest way to get subjects to equalize their responses is to give them feedback about any bias in their responses. With equal "yes" and "no" responses, the 75% correct corresponds to a z score of 0.674 and a $d' = 2z = 1.349$. If the subject does not have equal numbers of "yes" and "no" responses, then the d' would be calculated by $d' = z_{\text{hit}} - z_{\text{false alarm}}$. I do hope that Kaernbach's clever yes/no objective staircase will be explored by others. One should be able to enhance it with ratings and multiple stimuli (Klein, 2002).

Forced choice detection. As can be seen in the second column of Table 1, a popular category in this special issue is the forced choice method for detection. This method is used by Strasburger (2001b), Kaernbach (2001a), Linschoten et al. (2001), and Wichmann and Hill (2001a, 2001b). These researchers use PFs based on probability ordinates. The connection between P and d' is different from the yes/no case given by Equation 5. For 2AFC, signal detection theory provides a simple connection between d' and probability correct: $d'(x) = \sqrt{2}\, z(x)$. In the preceding section, I discussed the option of averaging the probabilities and then taking the z score (high threshold approach) or averaging the z scores and then calculating the probability (signal detection approach). The signal detection method is better, because it has a stronger empirical basis and avoids bias.

For an m-AFC task with $m > 2$, the connection between d' and P is more complicated than it is for $m = 2$. The connection, given by a fairly simple integral (Green & Swets, 1966; Macmillan & Creelman, 1991), has been tabulated by Hacker and Ratcliff (1979) and by Macmillan and Creelman (1991). One problem with these tables is that they are based on the assumption of unity ROC slope. It is known that there are many cases in which the ROC slope is not unity (Green & Swets, 1966), so these tables connecting

Table 1
Classification of the 10 Articles in This Special Issue According to Three Distinctions: Forced Choice Versus Yes/No, Detection Versus Discrimination, Adaptive Versus Constant Stimuli

	Detection		Discrimination		
Source	m-AFC	Yes/No	m-AFC	Yes/No	Adaptive or Constant Stimuli
Leek (2001)	General	loose γ	×	loose γ	A
Wichmann & Hill (2001a)	2AFC	(×)	(×)	(×)	C
Wichmann & Hill (2001b)	2AFC	(×)	(×)	(×)	C
Linschoten et al. (2001)	2AFC				A
Strasburger (2001a)	General	×	×		C
Strasburger (2001b)			10AFC		A
Kaernbach (2001a)	General		×		A
Kaernbach (2001b)	General	(1990)	×	(1990)	A
Miller & Ulrich (2001)		for $\gamma = 0$	× (for 0–100%)	×	C
Klein (present)	General	×	×	×	Both

d' and probability correct should be treated cautiously. W. P. Banks and I (unpublished) investigated this topic and found that near the $P = 50\%$ correct point, the dependence of d' on ROC slope is minimal. Away from the 50% point, the dependence can be strong.

Yes/No detection. Linschoten et al. (2001) compare three methods (limits, staircase, 2AFC likelihood) for measuring thresholds with a small number of trials. In the method of limits, one starts with a subthreshold stimulus and gradually increases the strength. On each trial, the observer says "yes" or "no" with respect to whether or not the stimulus is detected. Although this is a classic yes/no detection method, it will not be discussed in this article because there is no control of response bias. That is, blanks were not intermixed with signals.

In Table 1, the Wichmann and Hill (2001a, 2001b) articles are marked with ×s in parentheses because although these authors write only about 2AFC, they mention that all their methods are equally applicable for yes/no or *m*-AFC of any type (detection/discrimination). And their Matlab implementations are fully general. All the special issue articles reporting experimental data used a forced choice method. This bias in favor of the forced choice methodology is found not only in this special issue, it is widespread in the psychophysics community. Given the advantages of the yes/no method (see discussion in Section II), I hope that once yes/no adaptive methods are accepted, they will become the method of choice.

***m*-AFC discrimination.** It is common practice to represent 2AFC results as a plot of percent correct averaged over all trials versus stimulus strength. This PF goes from 50% to 100%. The asymmetry between the lower and upper asymptotes introduces some inefficiency in threshold estimation, as will be discussed. Another problem with the standard 50% to 100% plot is that a bias in choice of interval will produce an underestimate of d' as has been shown earlier. Researchers often use the 2AFC method because they believe that it avoids biased threshold estimates. It is therefore surprising that the relatively simple correction for 2AFC bias, discussed earlier, is rarely done.

The issue of bias in *m*-AFC also occurs when $m > 2$. In Strasburger's 10AFC letter discrimination task, it is common for subjects to have biases for responding with particular letters when guessing. Any imbalance in the response bias for different letters will result in a reduction of d' as it did in 2AFC.

There are several benefits of viewing the 2AFC discrimination data in terms of a PF going from 0% to 100%. Not only does it provide a simple way of viewing and calculating the interval bias, it also enables new methods for estimating the PF parameters such as those proposed by Miller and Ulrich (2001), as will be discussed in Section III. In my original comments on the Miller and Ulrich paper, I pointed out that because their nonparametric procedure has uniform weighting of the different PF levels, their method does not apply to 2AFC tasks. The asymmetric binomial error bars near the 50% and 100% levels cause the uniform weighting of the Miller and Ulrich approach to be nonoptimal. However, I now realize that the 2AFC discrimination task can be fit by a cumulative normal going from 0% to 100%. Because of that insight, I have marked the discrimination forced choice column of Table 1 for the Miller and Ulrich (2001) paper, with the proviso that the PF goes from 0 to 100%. Owing to the popularity of 2AFC, this modification greatly expands the relevance of their nonparametric approach.

Yes/No discrimination. Kaernbach (2001b) and Miller and Ulrich (2001) offer theoretical articles that examine properties of PFs that go from $P = 0\%$ to 100% ($P = p$ in Equation 1). In both cases, the PF is the cumulative normal (Equation 3). Although these PFs with $\gamma = 0$ could be for a yes/no detection task with a zero false alarm rate (not plausible) or a forced choice detection task with an infinite number of alternatives (not plausible either), I suspect that the authors had in mind a yes/no discrimination task (Table 1, col. 5). A typical discrimination task in vision is contrast discrimination, in which the observer responds to whether the presented contrast is greater than or less than a memorized reference. Feedback reinforces the stability of the reference. In a typical discrimination task, the reference is one exemplar from a continuum of stimulus strengths. If the reference is at a special zero point rather than being an element of a smooth continuum, the task is no longer a simple discrimination task. Zero contrast would be an example of a special reference. Klein (1985) discusses several examples which illustrate how a natural zero can complicate the analysis. One might wonder how to connect the PF from the detection regime in which the reference is zero contrast to the discrimination regime in which the reference (pedestal) is at a high contrast. The d' function to be introduced in Equation 20 does a reasonably good job of fitting data across the full range of pedestal strength, going from detection to discrimination.

The connection of the discrimination PF to the signal detection PF is the same as that for yes/no detection given in Equation 5: $d'(x) = z(x) - z(0)$, where z is the z score of the probability of a "greater than" judgment. In a discrimination task, the bias, $z(0)$, is the z score of the probability of saying "greater" when the reference stimulus is presented. The stimulus strength, x, that gives $z(x) = 0$ is the point of subjective equality (x = PSE). If the cumulative normal PF of Equation 3 is used, then the z score is linearly proportional to the stimulus strength $z(x) = (x - \mathrm{PSE})/\mathrm{threshold}$, where threshold is defined to be the point at which $d' = 1$. I will come back to these distinctions between detection and discrimination PFs after presenting more groundwork regarding thresholds, log abscissas, and PF shapes (see Equations 14 and 17).

Definition of Threshold

Threshold is often defined as the stimulus strength that produces a probability correct halfway up the PF. If humans operated according to a high-threshold assumption, this definition of threshold would be stable across different experimental methods. However, as I discussed following Equation 2, high-threshold theory has been discredited.

According to the more successful signal detection theory (Green & Swets, 1966), the d' at the midpoint of the PF changes according to the number of alternatives in a forced choice method and according to the false alarm rate in a yes/no method. This variability of d' with method is a good reason not to define threshold as the halfway point of the PF.

A definition of threshold that is relatively independent of the method used for its measurement is to define threshold as the stimulus strength that gives a fixed value of d'. The stimulus strength that gives $d' = 1$ (76% correct for 2AFC) is a common definition of threshold. Although I will show that higher d' levels have the advantage of giving more precise threshold estimates, unless otherwise stated I will take threshold to be at $d' = 1$ for simplicity. This definition applies to both yes/no and m-AFC tasks and to both detection and discrimination tasks.

As an example, consider the case shown in Figure 1, where the lower asymptote (false alarm rate) in a yes/no detection task is $\gamma = 15.87\%$, corresponding to a z score of $z = -1$. If threshold is defined to be at $d' = 1.0$, then, from Equation 5, the z score for threshold is $z = 0$, corresponding to a hit rate of 50% (not quite halfway up the PF). This example with a 50% hit rate corresponds to defining d' along the horizontal ROC axis. If threshold had been defined to be $d' = 2$ in Figure 1, then the probability correct at threshold would be 84.13%. This example, in which both the hit rate and correct rejection rate are equal (both are 84.13%), corresponds to the ROC negative diagonal.

Strasburger's Suggestion on Specifying Slope: A Logarithmic Abscissa?

In many of the articles in this special issue, a logarithmic abscissa such as decibels is used. Many shapes of PFs have been used with a log abscissa. Most delightful, but frustrating, is Strasburger's (2001a) paper on the PF maximum slope. He compares the Weibull, logistic, Quick, cumulative normal, hyperbolic tangent, and signal detection d' using a logarithmic abscissa. The present section is the outcome of my struggles with a number of issues raised by Strasburger (2001a) and my attempt to clarify them.

I will typically express stimulus strength, x, in threshold units,

$$x_t = \frac{x}{\alpha}, \tag{8}$$

where α is the threshold. Stimulus strength will be expressed in natural logarithmic units, y, as well as in linear units, x.

$$y_t = \log_e\left(x_t\right) = \log_e \frac{x}{\alpha} = y - Y, \tag{9}$$

where $y = \log_e(x)$ is the natural log of the stimulus and $Y = \log_e(\alpha)$ is the threshold on the log abscissa.

The slope of the psychometric function $P(y_t)$ with a logarithmic abscissa is

$$\text{slope}\left(y_t\right) = \frac{dP}{dy_t} \tag{10A}$$

$$= \left(1-\gamma\right)\frac{dp}{dy_t}, \tag{10B}$$

and the maximum slope is called β' by Strasburger (2001a). A very different definition of slope is sometimes used by psychophysicists, the log–log slope of the d' function, a slope that is constant at low stimulus strengths for many PFs. The log–log d' slope of the Weibull function is shown in the bottom pair of panels in Figure 1. The low-contrast log–log slope is β for a Weibull PF (Equation 12) and b for a d' PF (Equation 20). Strasburger (2001a) shows how the maximum slope using a probability ordinate is connected to the log–log slope, using a d' ordinate.

A frustrating aspect of Strasburger's article is that the slope units of P (probability correct per $\log_e$) are not familiar. Then it dawned on me that there is a simple connection between slope with a $\log_e$ abscissa, $\text{slope}_{\text{log}} = [dP(y_t)]/(dy_t)$, and slope with a linear abscissa, $\text{slope}_{\text{lin}} = [dP(x_t)]/(dx_t)$, namely:

$$\text{slope}_{\text{lin}} = \frac{dP\left(x_t\right)}{dx_t} = \frac{dP\left(y_t\right)}{dy_t}\cdot\frac{dy_t}{dx_t} = \frac{\text{slope}_{\text{log}}}{x_t}, \tag{11}$$

because $dy_t/dx_t = [d\log_e(x_t)]/(dx_t) = 1/x_t$. At threshold, $x_t = 1$ ($y_t = 0$), so at that point Strasburger's slope with a logarithmic axis is identical to my familiar slope plotted in threshold units on a linear axis. The simple connection between slope on the log and linear abscissas converted me to being a strong supporter of using a natural log abscissa.

The Weibull and cumulative normal psychometric functions. To provide a background for Strasburger's article, I will discuss three PFs: Weibull, cumulative normal, and d', as well as their close connections. A common parameterization of the PF is given by the Weibull function:

$$p_{\text{weib}}\left(x_t\right) = 1 - k^{x_t^{\beta}}, \tag{12}$$

where $p_{\text{weib}}(x_t)$, the PF that goes from 0% to 100%, is related to $P_{\text{weib}}(x_t)$, the probability of a correct response, by Equation 1; β is the slope; and k controls the definition of threshold. $p_{\text{weib}}(1) = 1-k$ is the percent correct at threshold ($x_t = 1$). One reason for the Weibull's popularity is that it does a good job of fitting actual data. In terms of logarithmic units, the Weibull function (Equation 12) becomes:

$$p_{\text{weib}}\left(y_t\right) = 1 - k^{\left[-\exp\left(\beta y_t\right)\right]}. \tag{13}$$

Panel a of Figure 1 is a plot of Equation 12 (Weibull function as a function of x on a linear abscissa) for the case $\beta = 2$ and $k = \exp(-1) = 0.368$. The choice $\beta = 2$ makes the Weibull an upside-down Gaussian. Panel d is the same function, this time plotted as a function of y corresponding to a natural log abscissa. With this choice of k, the point of maximum slope as a function of y_t is the threshold point ($y_t = 0$). The point of maximum slope, at $y_t = 0$, is marked with an asterisk in panel d. The same point in panel a at $x_t = 1$ is not the point of maximum slope on a linear abscissa, because of Equation 11. When plotted as a d' function [a z-score transform of $P(x_t)$] in panel b, the Weibull accelerates below threshold and decelerates above threshold, in agreement with a wide range of experimental data. The acceleration and deceleration are most clearly seen by the slope of d' in the log–log coordinates of panel e. The log–log d' slope is plotted in panels c and f. At $x_t = 0$, the log–log slope is 2, corresponding to our choice of $\beta = 2$. The slope falls surprisingly rapidly as stimulus strength increases, and the slope is near 1 at $x_t = 2$. If we had chosen $\beta = 4$ for the Weibull function, then the log–log d' slope would have gone from 4 at $x_t = 0$, to near 2 at $x_t = 2$. Equation 20 will present a d' function that captures this behavior.

Another PF commonly used is based on the cumulative normal (Equation 3):

$$p_\Phi\left(y_t\right) = \Phi\left(\frac{y_t}{\sigma}\right) \tag{14A}$$

or

$$z\left(y_t\right) = \frac{y_t}{\sigma} = \frac{y - Y}{\sigma}, \tag{14B}$$

where σ is the standard error of the underlying Gaussian function (its connection to β will be clarified later). Note that the cumulative normal PF in Equation 14 should be used only with a logarithmic abscissa, because y goes to $-\infty$, needed for the 0% lower asymptote, whereas the Weibull function can be used with both the linear (Equation 12) and the log (Equation 13) abscissa.

Two examples of Strasburger's maximum slope (his Equations 7 and 17) are

$$\beta' = \beta \exp(-1) = 0.368\beta \tag{15}$$

for the Weibull function (Equation 13), and

$$\beta' = \frac{1}{\sigma\sqrt{2\pi}} = \frac{0.399}{\sigma} \tag{16}$$

for the cumulative normal (Equation 14). In Equation 15, I set k in Equation 12 to be $k = \exp(-1)$. This choice amounts to defining threshold so that the maximum slope occurs at threshold ($y_t = 0$). Note that Equations 12–16 are missing the $(1-\gamma)$ factor that are present in Strasburger's Equations 7 and 17 because we are dealing with p rather than P. Equations 15 and 16 provide a clear, simple, and close connection between the slopes of the Weibull and cumulative normal functions, so I am grateful to Strasburger for that insight.

An example might help in showing how Equation 16 works. Consider a 2AFC task ($\gamma = 0.5$) assuming a cumulative normal PF with $\sigma = 1.0$. According to Equation 16, $\beta' = 0.399$. At threshold (assumed for now to be the point of maximum slope), $P(x_t = 1) = .75$. At 10% above threshold, $P(x_t = 1.1) \approx .75 + 0.1\,\beta' = 0.7899$, which is quite close to the exact value of .7896. This example shows how β', defined with a natural log abscissa, y_t, is relevant to a linear abscissa, x_t.

Now that the logarithmic abscissa has been introduced, this is a good place to stop and point out that the log abscissa is fine for detection but not for discrimination since the $x = 0$ point is often in the middle of the range. However, the cumulative normal PF is especially relevant to discrimination where the PF goes from 0% to 100% and would be written as

$$p_\Phi\left(x_t\right) = P_\Phi\left(x_t\right) = \Phi\left(\frac{x - \text{PSE}}{\alpha}\right) \tag{17A}$$

or

$$z_\Phi\left(x_t\right) = \frac{x - \text{PSE}}{\alpha}, \tag{17B}$$

where x = PSE is the point of subjective equality. The parameter, α, is the threshold, since $d'(\alpha) = z_\Phi(\alpha) - z_\Phi(0) = 1$. The threshold, α, is also σ standard deviations of the Gaussian probability density function (pdf). The reciprocal of α is the PF slope. I tend to use the letter σ for a unitless standard deviation, as occurs for the logarithmic variable, y. I use α for a standard deviation that has the units of the stimulus x, (like percent contrast). The comparison of Equation 14B for detection and Equation 17B for discrimination is useful. Although Equation 14 is used in most of the articles in this special issue, which are concerned with detection tasks, the techniques that I will be discussing are also relevant to discrimination tasks for which Equation 17 is used.

Threshold and the Weibull Function

To illustrate how a $d' = 1$ definition of threshold works for a Weibull function, let us start with a yes/no task in which the false alarm rate is $P(0) = 15.87\%$, corresponding to a z score of $z_{\text{FA}} = -1$, as in Figure 1. The z score at threshold ($d' = 1$) is $z_{\text{Th}} = z_{\text{FA}} + 1 = 0$, corresponding to a probability of $P(1) = 50\%$. From Equations 1 and 12, the k value for this definition of threshold is given by $k = [1-P(1)]/[1-P(0)] = 0.5943$. The Weibull function becomes

$$P_{\text{weib}}\left(x_t\right) = 1 - (1 - 0.1587) * 0.5943^{x_t^\beta}. \tag{18A}$$

If I had defined $d' = 2$ to be threshold then $z_{\text{Th}} = z_{\text{FA}} + 2 = 1$, leading to $k = (1-0.8413)/(1-0.1587) = 0.1886$, giving

$$P_{\text{weib}}\left(x_t\right) = 1 - (1 - 0.1587) * 0.1886^{x_t^\beta}. \tag{18B}$$

As another example, suppose the false alarm rate in a yes/no task is at 50%, not uncommon in a signal detection experiment with blanks and test stimuli intermixed. Then threshold at $d' = 1$ would occur at 84.13% and the PF would be

$$P_{\text{weib}}\left(x_t\right) = 1 - 0.5 * 0.3173^{x_t^\beta}, \tag{18C}$$

with $P_{\text{weib}}(0) = 50\%$ and $P_{\text{weib}}(1) = 84.13\%$. This case corresponds to defining d' on the ROC vertical intercept.

For 2AFC, the connection between d' and z is $z = d'/2^{0.5}$. Thus, $z_{\text{Th}} = 2^{-0.5} = 0.7071$, corresponding to $P_{\text{weib}}(1) = 76.02\%$, leading to $k = 0.4795$ in Equation 12. These connections will be clarified when an explicit form (Equation 20) is given for the d' function, $d'(x_t)$.

Complications With Strasburger's Advocacy of Maximum Slope

Here I will mention three complications implicated in Strasburger's suggestion of defining slope at the point of maximum slope on a logarithmic abscissa.

1. As can be seen in Equation 11, the slopes on linear and logarithmic abscissas are related by a factor of $1/x_t$. Because of this extra factor, the maximum slope occurs at a lower stimulus strength on linear as compared with log axes. This makes the notion of maximum slope less fundamental. However, since we are usually interested in the percent error of threshold estimates, it turns out that a logarithmic abscissa is the most relevant axis, supporting Strasburger's log abscissa definition.

2. The maximum slope is not necessarily at threshold (as Strasburger points out). For the Weibull functions defined in Equation 13 with $k = \exp(-1)$ and the cumulative normal function in Equation 14, the maximum slope (on a log axis) does occur at threshold. However, the two points are decoupled in the generalized Weibull function defined in Equation 12. For the Quick version of the Weibull function (Equation 13 with $k = 0.5$, placing threshold halfway up the PF), the threshold is below the point of maximum slope; the derivative of $P(x_t)$ at threshold is

$$\beta'_{\text{thresh}} = \frac{dP(x_t)}{dx_t} = (1-\gamma)\beta\frac{\log_e(2)}{2} = .347(1-\gamma)\beta,$$

which is slightly different from the maximum slope as given by Equation 15. Similarly, when threshold is defined at $d' = 1$, the threshold is not at the point of maximum slope. People who fit psychometric functions would probably prefer reporting slopes at the detection threshold rather than at the point of maximum slope.

3. One of the most important considerations in selecting a point at which to measure threshold is the question of how to minimize the bias and standard error of the threshold estimate. The goal of adaptive methods is to place trials at one level. In order to avoid a threshold bias due to a improper estimate of PF slope, the test level should be at the defined threshold. The variance of the threshold estimate when the data are concentrated as a single level (as with adaptive procedures) is given by Gourevitch and Galanter (1967) as

$$\text{var}(Y) = \frac{P(1-P)}{N}\left[\frac{dP(y_t)}{dy_t}\right]^{-2} \qquad (19)$$

If the binomial error factor of $P(1-P)/N$ were not present, the optimal placement of trials for measuring threshold would be at the point of maximum slope. However, the presence of the $P(1-P)$ factor shifts the optimal point to a higher level. Wichmann and Hill (2001b) consider how trial placement affects threshold variance for the method of constant stimuli. This topic will be considered in Section II.

Connecting the Weibull and d' Functions

Figures 5–7 of Strasburger (2001a) compare the log–log d' slope, b, with the Weibull PF slope, β. Strasburger's connection of $b = 0.88\beta$ is problematic, since the d' version of the Weibull PF does not have a fixed log–log slope. An improved d' representation of the Weibull function is the topic of the present section. A useful parameterization for d', which I shall refer to as the Stromeyer–Foley function, was introduced by Stromeyer and Klein (1974) and used extensively by Foley (1994):

$$d'(x_t) = \frac{x_t^b}{a + (1-a)x_t^{b+w-1}}, \qquad (20)$$

where x_t is the stimulus strength in threshold units as in Equation 8. The factors with a in the denominator are present so that d' equals unity at threshold ($x_t = 1$ or $x = \alpha$). At low x_t, $d' \approx x_t^b/a$. The exponent b (the log–log slope of d' at very low contrasts) controls the amount of facilitation near threshold. At high x_t, Equation 20 becomes $d' \approx x_t^{1-w}/(1-a)$. The parameter w is the log– log slope of the test threshold versus pedestal contrast function (the tvc or "dipper" function) at strong pedestals. One must be a bit cautious, however, because in yes/no procedures the tvc slope can be decoupled from w if the signal detection ROC curve has nonunity slope (Stromeyer & Klein, 1974). The parameter a controls the point at which the d' function begins to saturate. Typical values of these unitless parameters are $b = 2$, $w = 0.5$ and $a = 0.6$ (Yu, Klein, & Levi, 2001). The function in Equation 20 accelerates at low contrast (log–log slope of b) and decelerates at high contrasts (log–log slope of $1 - w$), in general agreement with a broad range of data.

For 2AFC, $z = d'/\sqrt{2}$, so from Equation 3, the connection between d' and probability correct is

$$P(x_t) = .5 + .5\,\text{erf}\left[\frac{d'(x_t)}{2}\right]. \qquad (21)$$

To establish the connection between the PFs specified in Equations 12 and 20–21, one must first have the two agree at threshold. For 2AFC with the $d' = 1$, the threshold at $P = .7602$ can be enforced by choosing $k = 0.4795$ (see the discussion following Equation 18C) so that Equations 1 and 12 become:

$$P(x_t) = 1 - (1-.5).4795^{x_t^\beta}. \qquad (22)$$

Modifying k as in Equation 22 leaves the Weibull shape unchanged and shifts only the definition of threshold. If $b = 1.06\beta$, $w = 1-0.39\beta$, and $a = 0.614$, then for *all* values of β, the Weibull and d' functions (Equations 21 and 22, vs. Equation 12) differ by less than 0.0004 for *all* stimulus strengths. At very low stimulus strengths, $b = \beta$. The value $b = 1.06\beta$ is a compromise for getting an overall optimal fit.

Strasburger (2001a) is concerned with the same issue. In his Table 1, he reports d' log–log slopes of [.8847

1.8379 3.131 4.421] for $\beta = [1\ 2\ 3.5\ 5]$. If the d' slope is taken to be a measure of the d' exponent b, then the ratio b/β is [0.8847 0.9190 0.8946 0.8842] for the four β values. Our value of $b/\beta = 1.06$ differs substantially from 0.8 (Pelli, 1985) and 0.88 (Strasburger, 2001a). Pelli (1985) and Strasburger (2001a) used d' functions with $a = 1$ in Equation 20. With $a = 0.614$, our d' function (Equation 20) starts saturating near threshold, in agreement with experimental data. The saturation lowers the effective value of b. I was very surprised to find that the Stromeyer–Foley d' function did such a good job in matching the Weibull function across the whole range of β and x. For a long time I had thought a different fit would be needed for each value of β.

For the same Weibull function in a yes/no method (false alarm rate = 50%), the parameters b and w are identical to the 2AFC case above. Only the value of a shifts from $a = 0.614$ (2AFC) to $a = 0.54$ (yes/no). In order to make $x_t = 1$ at $d' = 1$, k becomes 0.3173 (see Equation 18C) instead of 0.4795 (2AFC). As with 2AFC, the difference between the Weibull and d' function is less than .0004 (<.04%) for all stimulus strengths and all values of β and for both a linear and a logarithmic abscissa. These values mean that for both the yes/no and the 2AFC methods, the d' function (Equation 20) corresponding to a Weibull function with $\beta = 2$ has a low-contrast log–log slope of $b = 2.12$ and a high-contrast log–log slope of $1-w = 0.78$. The high-contrast tvc (test contrast vs. pedestal contrast discrimination function, sometimes called the "dipper" function) log–log slope would be $w = 0.22$.

I also carried out a similar analysis asking what cumulative normal σ is best matched to the Weibull. I found $\sigma = 1.1720/\beta$. However, the match is poor, with errors of >4% (rather than <0.04% in the preceding analysis). The poor fit of the two functions makes the fitting procedure sensitive to the weighting used in the fitting procedure. If one uses a weighting factor of $[P(1-P)/N]$, where P is the PF, one will find a quite different value for σ than if one ignored the weighting or if one chose the weighting on the basis of the data rather than the fitting function.

II. COMPARING EXPERIMENTAL TECHNIQUES FOR MEASURING THE PSYCHOMETRIC FUNCTION

This section, inspired by Linschoten et al. (2001), is an examination of various experimental methods for gathering data to estimate thresholds. Linschoten et al. compare three 2AFC methods for measuring smell and taste thresholds: a maximum-likelihood adaptive method, an up–down staircase method, and an ascending method of limits. The special aspect of their experiments was that in measuring smell and taste each 2AFC trial takes between 40 and 60 sec (Lew Harvey, personal communication) because of the long duration needed for the nostrils or mouth to recover their sensitivity. The purpose of the Linschoten et al. paper is to examine which of the three methods is most accurate when the number of trials is low. Linschoten et al. conclude that the 2AFC method is a good method for this task. My prior experience with forced choice tasks and low number of trials (Manny & Klein, 1985; McKee et al., 1985) made me wonder whether there might be better methods to use in circumstances in which the number of trials is limited. This topic is considered in Section II.

Compare 2AFC to Yes/No

The commonly accepted framework for comparing 2AFC and yes/no threshold estimates is signal detection theory. Initially, I will make the common assumption that the d' function is a power function (Equation 20, with $a = 1$):

$$d' = c_t^b = \exp\left(by_t\right). \tag{23}$$

Later in this section, the experimentally more accurate form, Equation 20, with $a < 1$, will be used. The variance of the threshold estimate will now be calculated for 2AFC and yes/no for the simplest case of trials placed at a single stimulus level, y, the goal of most adaptive methods. The PF slope relates ordinate variance to abscissa variance (from Gourevitch & Galanter, 1967):

$$\operatorname{var}(Y) = \frac{\operatorname{var}(d')}{\left(\dfrac{dd'}{dy_t}\right)^2}, \tag{24}$$

where $Y = y - y_t$ is the threshold estimate for a natural log abscissa as given in Equation 9. This formula for var(Y) is called the *asymptotic formula*, in that it becomes exact in the limit of large number of total trials, N. Wichmann and Hill (2001b) show that for the method of constant stimuli, a proper choice of testing levels brings one to the asymptotic region for $N < 120$ (the smallest value that they explored). Improper choice of testing levels requires a larger N in order to get to the asymptotic region. Equation 24 is based on having the data at a single level, as occurs asymptotically in adaptive procedures. In addition, the definition of threshold must be at the point being tested. If levels are tested away from threshold, there will be an additional contribution to the threshold variance (Finney, 1971; McKee et al., 1985) if the slope is a variable parameter in the fitting procedure. Equation 24 reaches its asymptotic value more rapidly for adaptive methods with focused placement of trials than with the constant stimulus method when slope is a free parameter.

Equation 24 needs analytic expressions for the derivative and the variance of d'. The derivative is obtained from Equation 23:

$$\frac{dd'}{dy_t} = bd'. \tag{25}$$

The variance of d' for both yes/no ($d' = z_{\text{hit}} + z_{\text{correct rejection}}$) and 2AFC ($d' = \sqrt{2}z_{\text{ave}}$) is

$$\operatorname{var}\left(d'\right) = 2\operatorname{var}(z) = 4\pi\exp\left(z^2\right)\operatorname{var}(P). \tag{26}$$

For the yes/no case, Equation 26 is based on a unity slope ROC curve with the subject's criterion at the negative diagonal of the ROC curve where the hit rate equals the correct rejection rate. This would be the most efficient operating point. The variance for a nonunity slope ROC curve was considered by Klein, Stromeyer, and Ganz (1974). From binomial statistics,

$$\mathrm{var}(P) = \frac{P(1-P)}{n}, \tag{27}$$

with $n = N$ for 2AFC and $n = N/2$ for yes/no since for this binary yes/no task the number of signal and blank trials is half the total number of trials.

Putting all of these factors together gives

$$\mathrm{var}(Y) = 4\pi \exp\left(z^2\right) \frac{P(1-P)}{n(bd')^2} \tag{28}$$

for both the yes/no and the 2AFC cases. The only difference between the two cases is the definition of n and the definition of d' ($d'_{\mathrm{2AFC}} = 2^{0.5}z$ and $d'_{\mathrm{YN}} = 2z$ for a symmetric criterion). When this is expressed in terms of N (number of trials) and z (z score of probability correct), we get

$$\mathrm{var}(Y) = 2\pi \exp\left(z^2\right) \frac{P(1-P)}{N(bz)^2} \tag{29}$$

for both 2AFC and yes/no. That is, when the percent corrects are all equal ($P_{\mathrm{2AFC}} = P_{\mathrm{hit}} = P_{\mathrm{correct\ rejection}}$), the threshold variance for 2AFC and yes/no are equal. The minimum value of var(Y) is $1.64/(Nb^2)$, occurring at $z = 1.57$ ($P = 94\%$). This value of 94% correct is the optimum value at which to test for both 2AFC and yes/no tasks, independent of b.

This result, that the threshold variance is the same for 2AFC and yes/no methods, seems to disagree with Figure 4 of McKee et al. (1985), where the standard errors of the threshold estimates are plotted for 2AFC and yes/no as a function of the number of trials. The 2AFC standard error (their Figure 4) is twice that of yes/no, corresponding to a factor of four in variance. However, the McKee et al. analysis is misleading, since all they (we) did for the yes/no task was to use Equation 2 to expand the PF to a 0 to 100% range, thereby doubling the PF slope. That rescaling is not the way to convert a 2AFC detection task to the yes/no task of the same detectability. A signal detection methodology is needed for that conversion, as was done in the present analysis.

One surprise is that the optimal testing probability is independent of the PF slope, b. This finding is a general property of PFs in which the dependence on the slope parameter has the form x^b. The 94% level corresponds to $d'_{\mathrm{YN}} = 3.15$ and $d'_{\mathrm{2AFC}} = 2.23$. For the commonly used $P = 75\%$ ($z = 0.765$, $d'_{\mathrm{YN}} = 1.35$, $d'_{\mathrm{2AFC}} = 0.954$) var($Y$) = $4.08/(Nb^2)$, which is about 2.5 times the optimal value, obtained by placing the stimuli at 94%. Green (1990) had previously pointed out that the 94% and 92% levels were the optimal points to test for the cumulative normal and logit functions, respectively, because of the minimum threshold estimate variability when one is testing at that point on the PF. Other PFs can have optimal points at slightly lower levels, but these are still above the levels to which adaptive methods are normally directed.

It is also useful to compare the 2AFC and yes/no at the same d' rather than at their optimal points. In that case, for d' values fixed at [1, 2, 3], the ratio of the 2AFC variance to the yes/no variance is [1.82, 1.36, 0.79]. At low d', there is an advantage for the 2AFC method, and that reverses at high d' as is expected from the optimal d' being higher for yes/no than for 2AFC. In general, one should run the yes/no method at d' levels above 2.0 (84% correct in the blank and signal intervals corresponds to $d' = 2$).

The equality of the optimal var(Y) for 2AFC and yes/no methods is not only surprising, it seems paradoxical—as can be seen by the following *gedanken* experiment. Suppose subjects close their eyes on the first interval of the 2AFC trial and base their judgments on just the second interval, as if they were doing a yes/no task, since blanks and stimuli are randomly intermixed. Instead of saying "yes" or "no," they would respond "Interval 2" or "Interval 1," respectively, so that the 2AFC task would become a yes/no task. The paradox is that one would expect the variance to be worse than for the eyes open 2AFC case, since one is ignoring information. However, Equation 29 shows that the 2AFC and yes/no methods have identical optimal variance. I suspect that the resolution of this paradox is that in the yes/no method with a stable criterion, the optimal variance occurs at a higher d' value (3.15) than the 2AFC value (2.23). The d' power law function that I chose does not saturate at large d' levels, giving a benefit to the yes/no method.

In order to check whether the paradox still holds with a more realistic PF (a d' function that saturates at large d'), I now compare the 2AFC and yes/no variance using a Weibull function with a slope β. In order to connect 2AFC and yes/no I need to use signal detection theory and the Stromeyer–Foley d' function (Equation 20) with the "Weibull parameters" $b = 1.06\beta$, $a = 0.614$, and $w = 1-.39\beta$. Following Equation 22, I pointed out that with these parameter values, the difference between the 2AFC Weibull function and the 2AFC d' function (converted to a probability ordinate) is less than .0004. After the derivative, Equation 25, is appropriately modified, we find that the optimal 2AFC variance is $3.42/(Nb^2)$. For the yes/no case, the optimal variance is $4.02/(Nb^2)$. This result resolves the paradox. The optimal 2AFC variance is about 18% lower than the yes/no variance, so closing one's eyes in one of the intervals does indeed hurt.

We shouldn't forget that if one counts stimulus presentations N_{p} rather than trials, N, the optimal 2AFC variance becomes $6.84/(N_{\mathrm{p}}b^2)$, as opposed to $4.02/(N_{\mathrm{p}}b^2)$ for yes/no. Thus for the Linschoten experiments with each stimulus presentation being slow, the yes/no methodology is more accurate for a given number of stimulus presentations, even at low d' values. Kaernbach (1990) compares 2AFC and yes/no adaptive methods with simulations and actual experiments. His abscissa is number of presentations, and his results are compatible with our findings.

The objective yes/no method that I have been discussing is inefficient, since 50% of the trials are blanks. Greater efficiency is possible if several points on the PF are measured (method of constant stimuli) with the number of blank trials the same as at the other levels. For example, if five levels of stimuli were measured, including blanks, then the blanks would be only 20% of the total rather than 50%. For the 2AFC paradigm, the subject would still have to look at 50% of the presentations being blanks. The yes/no efficiency can be further improved by having the subject respond to the multiple stimuli with multiple ratings rather than a binary yes/no response. This rating scale, method of constant stimuli, is my favorite method, and I have been using it for 20 years (Klein & Stromeyer, 1980). The multiple ratings allow one to measure the psychometric function shape for *d*′s as large as 4 (well into the dipper region where stimuli are slightly above threshold). It also allows one to measure the ROC slope (presence of multiplicative noise). Simulations of its benefits will be presented in Klein (2002).

Insight into how the number of responses affects threshold variance can be obtained by using the cumulative normal PF (Equation 17) considered by Kaernbach (2001b) and Miller and Ulrich (2001) with $\gamma = 0$. Typically, this is the case of yes/no discrimination, but it can also be 2AFC discrimination with an ordinate going from 0% to 100%, as I have discussed in connection with Figure 3. For a cumulative normal PF with a binary response and for a single stimulus level being tested, the threshold variance is (Equation 19) as follows:

$$\mathrm{var(thresh)} = \frac{\mathrm{var}(P)}{N\left(\dfrac{dP}{dy}\right)^2} = 2\pi P(1-P)\exp\left(z^2\right)\frac{\sigma^2}{N}, \quad (30)$$

from Equation 27 and forthcoming Equations 40 and 41. The optimal level to test is at $P = 0.5$, so the optimal variance becomes

$$\mathrm{var(thresh)} = \frac{\pi}{2}\cdot\frac{\sigma^2}{N}. \quad (31)$$

If instead of a binary response the observer gives an analog response (many, many rating categories), then the variance is the familiar connection between standard deviation and standard error:

$$\mathrm{var(thresh)} = \frac{\sigma^2}{N}. \quad (32)$$

With four response categories, the variance is $1.19\sigma^2/N$, which is closer to the analog than to the binary case. If stimuli with multiple strengths are intermixed (method of constant stimuli), then the benefit of going from a binary response to multiple responses is greater than that indicated in Equations 31–32 (Klein, 2002).

A brief list of other problems with the 2AFC method applied to detection comprises the following:

1. 2AFC requires the subject to memorize and then compare the results of two subjective responses. It is cognitively simpler to report each response as it arrives. Subjects without much experience can easily make mistakes (lapses), simply becoming confused about the order of the stimuli. Klein (1985) discusses memory load issues relevant to 2AFC.

2. The 2AFC methodology makes various types of analysis and modeling more difficult, because it requires that one average over all criteria. The response variable in 2AFC is the Interval 2 activation minus the Interval 1 activation (see the abscissa in Figure 2). This variable goes from minus infinity to plus infinity, whereas the yes/no variable goes from zero to infinity. It therefore seems more difficult to model the effects of probability summation and uncertainty for 2AFC than for yes/no.

3. Models that relate psychophysical performance to underlying processes require information about how the noise varies with signal strength (multiplicative noise). The rating scale method of constant stimuli not only measures *d*′ as a function of contrast, it also is able to provide an estimate of how the variance of the signal distribution (the ROC slope) increases with contrast. The 2AFC method lacks that information.

4. Both the yes/no and 2AFC methods are supposed to eliminate the effects of bias. Earlier I pointed out that in my recent 2AFC discrimination experiments, when the stimulus is near threshold, I find I have a strong bias in favor of responding "Stimulus 2." Until I learned to compensate for this bias (not typically done by naïve subjects) my *d*′ was underestimated. As I have discussed, it is possible to compensate for this bias, but that is rarely done.

Improving the Brownian (Up–Down) Staircase

Adaptive methods with a goal of placing trials at an optimal point on the PF are efficient. Leek (2001) provides a nice overview of these methods. There are two general categories of adaptive methods: those based on maximum (or mean) likelihood and those based on simpler staircase rules. The present section is concerned with the older staircase methods, which I will call Brownian methods. I used the name "Brownian" because of the up–down staircase's similarity to one-dimensional Brownian motion in which the direction of each step is random. The familiar up–down staircase is Brownian, because at each trial the decision about whether to go up or down is decided by a random factor governed by probability $P(x)$. I want to make this connection to Brownian motion, because I suspect that there are results in the physics literature that are relevant to our psychophysical staircases. I hope these connections get made.

A Brownian staircase has a minimal memory of the run history; it needs to remember the previous step (including direction) and the previous number of reversals or number of trials (used for when to stop and for when to decrease step size). A typical Brownian rule is to decrease signal strength by one step (like 1 dB) after a correct response and to increase it three steps after an incorrect response (a one down–three up rule). I was pleased to learn recently that his rule works nicely for an objective yes/no task (Kaernbach, 1990) as well as 2AFC, as discussed earlier.

In recent years, likelihood methods (Pentland, 1980; Watson & Pelli, 1983) have become popular and have somewhat displaced Brownian methods. There is a widespread feeling that likelihood methods have substantially smaller threshold variance than do staircases in which thresholds are obtained by simply averaging the levels tested. Given the importance of this issue for informing researchers about how best to measure thresholds, it is surprising that there have been so few simulations comparing staircases and optimal methods. Pentland's (1980) results, showing very large threshold variance for a staircase method, are so dramatically different from the results of my simulations that I am skeptical. The best previous simulations of which I am aware are those of Green (1990), who compared the threshold variance from several 2AFC staircase methods with the ideal variance (Equation 24). He found that as long as the staircase rule placed trials above 80% correct, the ratio of observed to ideal threshold variance was about 1.4. This value is the square of the inverse of Green's finding that the ratio of the expected to the observed sigma is about .85.

In my own simulations (Klein, 2002), I found that as long as the final step size was less than 0.2σ, the threshold variance from simply averaging over levels was close to the optimal value. Thus, the staircase threshold variance is about the same as the variance found by other optimal methods. One especially interesting finding was that the common method of averaging an even number of reversal points gave a variance that was about 10% higher than averaging over all levels. This result made me wonder how the common practice of averaging reversals ever got started. Since Green (1990) averaged over just the reversal points, I suspect that his results are an overestimate of the staircase variance. In addition, since Green specifies step size in decibels, it is difficult for me to determine whether his final step size was small enough. Since the threshold variance is inversely related to slope, it is important to relate step size to slope rather than to decibels. The most natural reference unit is σ, the standard deviation associated with the cumulative normal. In summary, I suggest that this general feeling of distrust of the simple Brownian staircase is misguided, and that simple staircases are often as good as likelihood methods. In the section Summary of Advantages of Brownian Staircase, I point out several advantages of the simple staircase over the likelihood methods.

A qualification should be made to the claim above that averaging levels of a simple staircase is as good as a fancy likelihood method. If the staircase method uses a too small step size at the beginning and if the starting point is far from threshold, then the simple staircase can be inefficient in reaching the threshold region. However, at the end of that run, the astute experimenter should notice the large discrepancy between the starting point and the threshold, and the run should be redone with an improved starting point or with a larger initial step size that will be reduced as the staircase proceeds.

Increase number of 2AFC response categories. Independent of the type of adaptive method used, the 2AFC task has a flaw whereby a series of lucky guesses at low stimulus levels can erroneously drive adaptive methods to levels that are too low. Some adaptive methods with a small number of trials are unable to fully recover from a string of lucky guesses early in the run. One solution is to allow the subject to have more than two responses. For reasons that I have never understood, people like to have only two response categories in a 2AFC task. Just as I am reluctant to take part in 2AFC experiments, I am even more reluctant to take part in two response experiments. The reason is simple. I always feel that I have within me more than one bit of information after looking at a stimulus. I usually have at least four or more subjective states (2 bits) for a yes/no task: HN, LN, LY, HY, where H and L are *high* and *low confidence* and N and Y are *did not* and *did see it*. For a 2AFC, my internal states are H1, L1, L2, H2 for high and low confidence on Intervals 1 and 2. Kaernbach (2001a) in this special issue does an excellent job of justifying the low-confidence response category. He provides solid arguments, simulations, and experimental data on how a low-confidence category can minimize the "lucky guess" problem.

Kaernbach (2001a) groups the L1 and L2 categories together into a "don't know" category and argues persuasively that the inclusion of this third response can increase the precision and accuracy of threshold estimates, while also leaving the subject a happier person. He calls it the "unforced choice" method. Most of his article concerns the understandable fear of introducing a response bias, exactly what 2AFC was invented to avoid. He has developed an intelligent rule for what stimulus level should follow a "don't know" response. Kaernbach shows with both simulations and real data that with his method this potential response bias problem has only a second-order effect on threshold estimates (similar to the 2AFC interval bias discussed around Equation 7), and its advantages outweigh the disadvantages. I urge anyone using the 2AFC method to read it. I conjecture that Kaernbach's method is especially useful in trimming the lower tail of the distribution of threshold estimates and also in reducing the interval bias of 2AFC tasks since the low confidence rating would tend to be used for trials on which there is the most uncertainty.

Although Kaernbach has managed to convince me of the usefulness of his three-response method, he might have trouble convincing others because of the response bias question. I should like to offer a solution that might quell the response bias fears. Rather than using 2AFC with three responses, one can increase the number of responses to four—H1, L1, L2, H2—as discussed two paragraphs above. The L rating can be used when subjects feel they are guessing.

To illustrate how this staircase would work, consider the case in which we would like the 2AFC staircase to converge to about 84% correct. A 5 up–1 down staircase (5 levels up for a wrong response, and 1 level down for a correct response) leads to an average probability correct of 5/6 = 83.33%. If one had zero information about the trial and guessed, then one would be correct half the time and the staircase would shift an average of (5−1)/2 = +2 levels. Thus, in his scheme with a single "don't know" response, Kaernbach would have the staircase move up 2 levels following a "don't know" response. One might worry that continued use of the "don't know" response would continuously increase the level so that one would get an upward bias. But Kaernbach points out that the "don't know" response replaces incorrect responses to a greater extent than it replaces correct responses, so that the +2 level shift is actually a more negative shift than the +5 shift associated with a wrong response. For my suggestion of four responses, the step shifts would be –1, +1, +3, and +5 levels for responses HC, LC, LI, and HI, where H and L stand for high and low confidence and C and I stand for correct and incorrect. A slightly more conservative set of responses would be –1, 0, +4, +5, in which case the low-confidence correct judgment leaves the level unchanged. In all these examples, including Kaernbach's three-response method, the low-confidence response has an average level shift of +2 when one is guessing. The most important feature of the low confidence rating is that when one is below threshold, giving a low confidence rating will prevent one from long strings of lucky guesses that bring one to an inappropriately low level. This is the situation that can be difficult to overcome if it occurs in the early phase of a staircase when the step size can be large. The use of the confidence rating would be especially important for short runs, where the extra bit of information from the subject is useful for minimizing erroneous estimates of threshold.

Another reason not to be fearful of the multiple-response 2AFC method is that after the run is finished, one can estimate threshold by using a maximum likelihood procedure rather than by averaging the run locations. Standard likelihood analysis would have to be modified for Kaernbach's three-response method in order to deal with the "don't know" category. For the suggested four-response method, one could simply ignore the confidence ratings for the likelihood analysis. My simulations, discussed earlier in this section, and Green's (1990) reveal that averaging over the levels (excluding a number of initial levels to avoid the starting point bias) has about as good a precision and accuracy as the best of the likelihood methods. If the threshold estimates from the likelihood method and from the averaging levels method disagree, that run could be thrown out.

Summary of advantages of Brownian staircase. Since there is no substantial variance advantage to the fancier likelihood methods, one can pay more attention to a number of advantages associated with simple staircases.

1. At the beginning of a run, likelihood methods may have too little inertia (the sequential stimulus levels jump around easily), unless a "stiff" prior is carefully chosen. By *inertia*, I refer to how difficult it is for the next level to deviate from the present level. At the beginning of a run, one would like to familiarize the subject with the stimulus by presenting a series of stimuli that gradually increase in difficulty. This natural feature of the Brownian staircase is usually missing in QUEST-like methods. The slightly inefficient first few trials of a simple staircase may actually be a benefit, since it gives the subject some practice with slightly visible stimuli.

2. Toward the end of a maximum likelihood run, one can have the opposite problem of excessive inertia, whereby it is difficult to have substantial shifts of test contrast. This can be a problem if nonstationary effects are present. For example, thresholds can increase in mid-run, owing to fatigue or to adaptation if a suprathreshold pedestal or mask is present. In an old-fashioned Brownian staircase, with moderate inertia, a quick look at the staircase track shows the change of threshold in mid-run. That observation can provide an important clue to possible problems with the current procedures. A method that reveals nonstationarity is especially important for difficult subjects, such as infants or patients, in whom fatigue or inattention can be an important factor.

3. Several researchers have told me that the simplicity of the staircase rules is itself an advantage. It makes writing programs easier. It is nicer for teaching students. It also simplifies the uncertain business of guessing likelihood priors and communicating the methodology when writing articles.

Methods That Can Be Used to Estimate Slope as Well as Threshold

There are good reasons why one might want to learn more about the PF than just the single threshold value. Estimating the PF slope is the first step in this direction. To estimate slope, one must test the PF at more than one point. Methods are available to do this (Kaernbach, 2001b; Kontsevich & Tyler, 1999), and one could even adapt simple Brownian staircase procedures to give better slope estimates with minimum harm to threshold estimates (see the comments in Strasburger, 2001b).

Probably the best method for getting both thresholds and slopes is the Ψ (Psi) method of Kontsevich and Tyler (1999). The Ψ method is not totally novel, it borrows techniques from many previous researchers—as Kontsevich and Tyler point out in a delightful paragraph that both clarifies their method and states its ancestry. This paragraph summarizes what is probably the most accurate and precise approach to estimating thresholds and slope, so I reproduce it here:

> To summarize, the Ψ method is a combination of solutions known from the literature. The method updates the posterior probability distribution across the sampled space of the psychometric functions based on Bayes' rule (Hall, 1968; Watson & Pelli, 1983). The space of the psychometric functions is two-dimensional (King-Smith & Rose, 1997; Watson & Pelli, 1983). Evaluation of the psychometric function is based on computing the mean of the posterior probability distribution (Emerson, 1986; King-Smith et al., 1994). The termination rule is based on the number of trials, as the most practical option (King-Smith

et al., 1994; Watson & Pelli, 1983). The placement of each new trial is based on one-step ahead minimum search (King-Smith, 1984) of the expected entropy cost function (Pelli, 1987).

A simplified, rough description of the Ψ trial placement for 2AFC runs is that initially trials are placed at about the 90% correct level to establish a solid threshold. Once threshold has been roughly established, trials are placed at about the 90% and the 70% levels to pin down the slope. The precise levels depend on the assumed PF shape.

Before one starts measuring slopes, however, one must first be convinced that knowledge of slope can be important. It is so, for several reasons:

1. Parametric fitting procedures either require knowledge of slope, β, or must have data that adequately constrain the slope estimate. Part of the discussion that follows will be on how to obtain reliable threshold estimates when slope is not well known. A tight knowledge of slope can minimize the error of the threshold estimate.

2. Strasburger (2001a, 2001b) gives good arguments for the desirability of knowing the shape of the PF. He is interested in differences between foveal and peripheral visual processing, and he uses the PF slope as an indication of differences.

3. Slopes are different for different stimuli, and this can lead to misleading results if slope is ignored. It may help to describe my experience with the Modelfest project (Carney et al., 2000) to make this issue concrete. This continuing project involves a dozen laboratories around the world. We have measured detection thresholds for a wide variety of visual stimuli. The data are used for developing an improved model of spatial vision. Some of the stimuli were simple, easily memorized low spatial frequency patterns that tend to have shallower slopes because a stimulus-known-exactly template can be used efficiently and with minimal uncertainty. On the other hand, tiny Modelfest stimuli can have a good deal of spatial uncertainty and are therefore expected to produce PFs with steeper slopes. Because of the slope change, the pattern of thresholds will depend on the level at which threshold is measured. When this rich dataset is analyzed with filter models that make estimates for mechanism bandwidths and mechanism pooling, these different slope-dependent threshold patterns will lead to different estimates for the model parameters. Several of us in the Modelfest data gathering group argued that we should use a data-gathering methodology that would allow estimation of PF slope of each of the 45 stimuli. However, I must admit that when I was a subject, my patience wore thin and I too chose the quicker runs. Although I fully understand the attraction of short runs focused on thresholds and ignoring slope, I still think there are good reasons for spreading trials out. For one thing, interspersing trials that are slightly visible helps the subject maintain memory of the test pattern.

4. Slopes can directly resolve differences between models. In my own research, for example, my colleagues and I are able to use the PF slope to distinguish long-range facilitation that produces a low-level excitation or noise reduction (typical with a cross-orientation surround) from a higher level uncertainty reduction (typical with a same-orientation surround).

Throwing-Out Rules, Goodness-of-Fit

There are occasions when a staircase goes sour. Most often, this happens when the subject's sensitivity changes in the middle of a run, possibly because of fatigue. It is therefore good to have a well-specified rule that allows nonstationary runs to be thrown out. The throwing-out rule could be based on whether the PF slope estimate is too low or too high. Alternatively, a standard approach is to look at the chi-square goodness-of-fit statistic, which would involve an assumption about the shape of the underlying PF. I have a commentary in Section III regarding biases in the chi-square metric that were found by Wichmann and Hill (2001a). Biases make it difficult to use standard chi-square tables. The trouble with this approach of basing the goodness-of-fit just on the cumulated data is that it ignores the sequential history of the run. The Brownian up–down staircase makes the sequential history vividly visible with a plot of sequentially tested levels. When fatigue or adaptation sets in, one sees the tested levels shift from one stable point to another. Leek, Hanna, and Marshall (1991) developed a method of assessing the nonstationarity of a psychometric function over the course of its measurement, using two interleaved tracks. Earlier, Hall (1983) also suggested a technique to address the same problem. I encourage further research into the development of a non-stationarity statistic that can be used to discard bad runs. This statistic would take the sequential history into account as well as the PF shape and the chi-square.

The development of a "throwing-out" rule is the extreme case of the notion of weighted averaging. One important reason for developing good estimators for goodness-of-fit and good estimators for threshold variance is to be able to optimally combine thresholds from repeated experiments. I shall return to this issue in connection with the Wichmann and Hill (2001a) analysis of bias in goodness-of-fit metrics.

Final Thought: The Method Should Depend on the Situation

Also important is that the method chosen should be appropriate for the task. For example, suppose one is using well-experienced observers for testing and retesting similar conditions in which one is looking for small changes in threshold or one is interested in the PF shape. In such cases, one has a good estimate of the expected threshold, and the signal detection method of constant stimuli with blanks and with ratings might be best. On the other hand, for clinical testing in which one is highly uncertain about an individual's threshold and criterion stability, a 2AFC likelihood or staircase method (with a confidence rating) might be best.

III. DATA-FITTING METHODS FOR ESTIMATING THRESHOLD AND SLOPE, PLUS GOODNESS-OF-FIT ESTIMATION

This final section of the paper is concerned with what to do with the data after they have been collected. We have here the standard Bayesian situation. Given the PF, it is easy to generate samples of data, with confidence intervals. But we have the inverse situation, in which we are given the data and need to estimate properties of the PF and get confidence limits for those estimates. The pair of articles in this special issue by Wichmann and Hill (2001a, 2001b), as well as the articles by King-Smith et al. (1994), Treutwein (1995), and Treutwein and Strasburger (1999) do an outstanding job of introducing the issues and provide many important insights. For example, Emerson (1986) and King-Smith et al. emphasize the advantages of mean likelihood over maximum likelihood. The speed of modern computers makes it just as easy to calculate mean likelihood as it is to calculate maximum likelihood. As another example, Treutwein and Strasburger (1999) demonstrate the power of using Bayesian priors for estimating parameters by showing how one can estimate all four parameters of Equation 1A in fewer than 100 trials. This important finding deserves further examination.

In this section, I will focus on four items. First I will discuss the relative merits of parametric versus nonparametric methods for estimating the PF properties. The paper by Miller and Ulrich (2001) provides a detailed analysis of a nonparametric method for obtaining all the moments of the PF. I will discuss some advantages and limitations of their method. Second is the question of goodness-of-fit. Wichmann and Hill (2001a) show that when the number of trials is small, the goodness-of-fit can be very biased. I will examine the cause of that bias. Third is the issue discussed by Wichmann and Hill (2001a), that of how lapses effect slope estimates. This topic is relevant to Strasburger's (2001b) data and analysis. Finally, there is the possibly controversial topic of an upward slope bias in adaptive methods that is relevant to the papers of Kaernbach (2001b) and Strasburger (2001b).

Parametric Versus Nonparametric Fits

The main split between methods for estimating parameters of the psychometric function is the distinction between parametric and nonparametric methods. In a parametric method, one uses a PF that involves several parameters and uses the data to estimate the parameters. For example, one might know the shape of the PF and estimate just the threshold parameter. McKee et al. (1985) show how slope uncertainty affects threshold uncertainty in probit analysis. The two papers by Wichmann and Hill (2001a, 2001b) are an excellent introduction to many issues in parametric fitting. An example of a nonparametric method for estimating threshold would be to average the levels of a Brownian staircase after excluding the first four or so reversals in order to avoid the starting level bias. Before I served as guest editor of this special symposium issue, I believed that if one had prior knowledge of the exact PF shape, a parametric method for estimating threshold was probably better than nonparametric methods. After reading the articles presented here, as well as carrying out my recent simulations, I no longer believe this. I think that, in the proper situation, nonparametric methods are as good as parametric methods. Furthermore, there are situations in which the shape of the PF is unknown and nonparametric methods can be better.

Miller and Ulrich's Article on the Nonparametric Spearman–Kärber Method of Analysis

The standard way to extract information from psychometric data is to assume a functional shape such as a Weibull, cumulative normal, logit, d', and then vary the parameters until likelihood is maximized or chi-square is minimized. Miller and Ulrich (2001) propose using the Spearman–Kärber (SK) nonparametric method, which does not require a functional shape. They start with a probability density function, defined by them as the derivative of the PF:

$$\mathrm{pdf}(i.) = \frac{P(i) - P(i-1)}{s(i) - s(i-1)}, \tag{33}$$

where $s(i)$ is the stimulus intensity at the ith level, and $P(i)$ is the probability correct at the that level. The notation $\mathrm{pdf}(i.)$ is meant to indicate that it is the pdf for the interval between $i-1$ and i. Miller and Ulrich give the rth moment of this distribution as their Equation 3:

$$\mu_r = \frac{1}{r+1} \sum \mathrm{pdf}(i.)\left[s(i)^{r+1} - s(i-1)^{r+1}\right]. \tag{34}$$

The next four equations are my attempt to clarify their approach. I do this because, when applicable, it is a fine method that deserves more attention. Equation 34 probably came from the mean value theorem of calculus:

$$f\left[s(i)\right] - f\left[s(i-1)\right] = f'(s_{\mathrm{mid}})\left[s(i) - s(i-1)\right], \tag{35}$$

where $f(s) = s^{r+1}$, and $f'(s_{\mathrm{mid}}) = (r+1)s_{\mathrm{mid}}^{r}$ is the derivative of f with respect to s somewhere within the interval between $s(i-1)$ and $s(i)$. For a slowly varying function, s_{mid} would be near the midpoint. Given Equation 35, Equation 34 becomes

$$\mu_r = \sum_i \mathrm{pdf}(i.) s_{\mathrm{mid}}^{r} \Delta s \ , \tag{36}$$

where $\Delta s = s(i) - s(i-1)$. Equation 36 is what one would mean by the rth moment of the distribution. The case $r = 0$ is important, since it gives the area under the pdf,

$$\mu_0 = \sum_i \mathrm{pdf}(i.) \ \Delta s = P(i_{\max}) - P(i_{\min}), \tag{37}$$

by using the definition in Equation 33. The summation in Equation 37 gives the area under the pdf. For it to be a proper pdf, its area must be unity, which means that the probability at the highest level, $P(i_{\max})$, must be unity and the probability at the lowest level, $P_{\min}$, must be zero, as Miller and Ulrich (2001) point out.

The next simplest SK moment is for $r = 1$:

$$\mu_1 = \sum \text{pdf}(i.)\frac{s(i)^2 - s(i-1)^2}{2} = \sum\left[P(i) - P(i-1)\right]\frac{s(i) + s(i-1)}{2} \qquad (38)$$

from Equation 33. This first moment provides an estimate of the centroid of the distribution corresponding to the center of the psychometric function. It could be used as an estimate of threshold for detection or PSE for discrimination. An alternate estimator would be the median of the pdf, corresponding to the 50% point of the PF (the proper definition of PSE for discrimination).

Miller and Ulrich (2001) use Monte Carlo simulations to claim that the SK method does a better job of extracting properties of the psychometric function than do parametric methods. One must be a bit careful here, since in their parametric fitting they generate the data by using one PF and then fit it with other PFs. But still it is a surprising claim; if this simple method is so good, why have people not been using it for years? There are two possible answers to this question. One is that people never thought of using it before Miller and Ulrich. The other possibility is that it has problems. I have concluded that the answer is a combination of both factors. I think that there is an important place for the SK approach, but its limitations must be understood. That is my next topic.

The most important limitation of the SK nonparametric method is that it has only been shown to work well for method of constant stimulus data in which the PF goes from 0 to 1. I question whether it can be applied with the same confidence to 2AFC and to adaptive procedures with unequal numbers of trials at each level. The reason for this limitation is that the SK method does not offer a simple way to weight the different samples. Let us consider two situations with unequal weighting: adaptive methods and 2AFC detection tasks.

Adaptive methods place trials near threshold, but with very uneven distribution of number of trials at different levels. The SK method gives equal weighting to sparsely populated levels that can be highly variable, thus contributing to a greater variance of parameter estimates. Let us illustrate this point for an extreme case with five levels $s = [1\ 2\ 3\ 4\ 5]$. Suppose the adaptive method placed [1 1 1 98 1] trials at the five levels and the number correct were [0 1 1 49 1], giving probabilities [0 1. 1. .5 1.] and pdf = [1 0 −.5 .5]. The following PEST-like (Taylor & Creelman, 1967) rules could lead to this unusual data: Move down a level if the presently tested level has greater than P_1% correct, move up three levels if the presently tested level has less than P_2% correct. P_1 can be any number less than 33%, and P_2 can be any number greater than 67%. The example data would be produced by starting at Level 5 and getting four in a row correct followed by an error, followed by a wide set possible of alternating responses for which the PEST rules would leave the fourth level as the level to be tested. The SK estimate of the center of the PF is given by Equation 38 to be $\mu_1 = 2.00$. Since a pdf with a negative probability is distasteful to some, one can monotonize the PF according to the procedure of Miller and Ulrich (which does include the number of trials at each level). The new probabilities are [0 .51 .51 .51 1] with pdf = [.51 0 0 .49]. The new estimate of threshold is $\mu_1 = 2.97$. However, given the data with 49/98 correct at Level 4, an estimate that includes a weighting according to the number of trials at each level would place the centroid of the distribution very close to $\mu_1 = 4$. Thus we see that because the SK procedure ignores the number of trials at each level, it produces erroneous estimates of the PF properties. This example with shifting thresholds was chosen to dramatize the problems of unequal weighting and the effect of monotonizing.

A similar problem can occur even with the method of constant stimuli with equal number of trials at each level, but with unequal binomial weighting as occurs in 2AFC data. The unequal weighting occurs because of the 50% lower asymptote. The binomial statistics error bar of data at 55% correct is much larger than for data at 95% correct. Parametric procedures such as probit analysis give less weighting to the lower data. This is why the most efficient placement of trials in 2AFC is above 90% correct, as I have discussed in Section II. Thus it is expected that for 2AFC, the SK method would give threshold estimates less precise than the estimates given by a parametric method that includes weighting.

I originally thought that the SK method would fail on all types of 2AFC data. However, in the discussion following Equation 7, I pointed out that for 2AFC discrimination tasks, there is a way of extending the data to the 0–100% range that produces weightings similar to those for a yes/no task. Rather than use the standard Equation 2 to deal with the 50% lower asymptote, one should use the method discussed in connection with Figure 3, in which one can use for the ordinate the percent of correct responses in Interval 2 and use the signal strength in Interval 2 minus that in Interval 1 for the abscissa. That procedure produces a PF that goes from 0% to 100%, thereby making it suitable for SK analysis while also removing a possible bias.

Now let us examine the SK method as applied to data for which it is expected to work, such as discrimination PFs that go from 0 to 100%. These are tasks in which the location of the psychometric function is the PSE and the slope is the threshold.

A potentially important claim by Miller and Ulrich (2001) is that the SK method performs better than probit analysis. This is a surprising claim, since their data are generated by a probit PF; one would expect probit analysis to do well, especially since probit analysis does especially well with discrimination PFs that go from 0 to 100%. To help clarify this issue, Miller and I did a number of simulations. We focused on the case of $N = 40$, with 8 trials at each of 5 levels. The cumulative normal PF (Equation 14) was used with equally spaced z scores and with the lowest and highest probabilities being 1% and 99%. The probabilities at the sample points are $P(i) = 1\%, 12.22\%,$

50%, 87.78%, and 99%, corresponding to z scores of $z(i) = -2.328$, -1.164, 0, 1.164, and 2.328. Nonmonotonicity is eliminated as Miller and Ulrich suggest (see the Appendix for Matlab Code 4, for monotonizing). I did Monte Carlo simulations to obtain the standard deviation of the location of the PF, μ_1, given by Equation 38. I found that both the SK method and the parametric probit method (assuming a fixed unity slope) gave a standard deviation of between 0.28 and 0.29. Miller and Ulrich's result, reported in their Table 17, gives a standard error of 0.071. However, their Gaussian had a standard deviation of 0.25, whereas my z-score units correspond to a standard deviation of unity. Thus their standard error must be multiplied by 4, giving a standard error of 0.284, in excellent agreement with my value. So, at least in this instance, the SK method did not do better than the parametric method (with slope fixed), and the surprise mentioned at the beginning of this paragraph did not materialize. I did not examine other examples where nonparametric methods might do better than the matched parametric method. However, the simulations of McKee et al. (1985) give some insight into why probit analysis with floating slope can do poorly even on data generated from probit PFs. McKee et al. show that when the number of trials is low and if stimulus levels near both 0% and 100% are not tested, the slope can be very flat and the predicted threshold can be quite distant. If care is not taken to place a bound on these probit estimates, one would expect to find deviant threshold estimates. The SK method, on the other hand, has built-in bounds for threshold estimates, so outliers are avoided. These bounds could explain why SK can do better than probit. When the PF slope is uncertain, nonparametric methods might work better than parametric methods, a decided advantage for the SK method.

It is instructive to also do an analytic calculation of the variance based on an analysis similar to what was done earlier. If only the ith level of the five had been tested, the variance of the PF location (the PSE in a discrimination task) would have been as follows (Gourevitch & Galanter, 1967):

$$\mathrm{var(PSE)}_i = \frac{\mathrm{var}\left(P_i\right)}{\left(\dfrac{dP_i}{dx_i}\right)^2} \tag{39}$$

where var(P) is given by Equation 27 and PF slope is

$$\frac{dP_i}{dx_i} = \frac{dP_i}{dz_i} \cdot \frac{dz_i}{dx_i}, \tag{40}$$

where for the cumulative normal, $dz_i/dx_i = 1/\sigma$, where σ is the standard deviation of the pdf and

$$\left(\frac{dP_i}{dz_i}\right)^2 = \frac{\exp\left(-z_i^2\right)}{2\pi} \tag{41}$$

from Equation 3A. By combining these equations one gets

$$\mathrm{var(PSE)}_i = 2\pi\sigma^2 P_i\left(1-P_i\right)\frac{\exp\left(z_i^2\right)}{N_i}. \tag{42}$$

Note that if all trials were placed at the optimal stimulus location of $z = 0$ ($P = .5$), Equation 42 would become $\sigma^2\pi/2N_i$, as discussed earlier. When all five levels are combined, the total variance is smaller than each individual variance, as given by the variance summation formula:

$$\mathrm{var}_{\mathrm{tot}}^{-1} = \sum \mathrm{var}_i^{-1}. \tag{43}$$

Finally, upon plugging in the five values of z_i ranging from -2.328 to $+2.328$ and P_i ranging from 1% to 99%, the standard error of the PF location is

$$SE = \mathrm{var}_{\mathrm{tot}}^{1/2} = 0.2844. \tag{44}$$

This value is precisely the value obtained in our Monte Carlo simulations of the nonparametric SK method and also by the slope fixed parametric probit analysis. This triple confirmation shows that the SK method is excellent in the realm where it applies (equal number of trials at levels that span 0 to 100%). It also confirms our simple analytic formula.

The data presented in Miller and Ulrich's Table 17 give us a fine opportunity to ask about the efficiency of the method of constant stimuli that they use. For the $N = 40$ example discussed here, the variance is

$$\mathrm{var}_{\mathrm{tot}} = SE^2 = \frac{3.24}{N_{\mathrm{tot}}}. \tag{45}$$

This value is more than twice the variance of $\pi/(2N_{\mathrm{tot}})$ that is obtained for the same PF using adaptive methods, including the simple up–down staircase, that place trials near the optimal point.

The large variance for the Miller and Ulrich stimulus placement is not surprising, given that, of the five stimulus levels, the two at $P = 1\%$ and 99% are quite inefficient. The inefficient placement of trials is a common problem for the method of constant stimuli as opposed to adaptive methods. However, as I mentioned earlier, the method of constant stimuli combined with multiple response ratings in a signal detection framework may be the best of all methods for detection studies for revealing properties of the system near threshold while maintaining a high efficiency.

A curious question comes up about the optimal placement of trials for measuring the location of the psychometric function. With the original function, the optimal placement is at the steepest part of the function (at $P = 0$). However, with the SK method, when one is determining the location of the pdf one might think that the optimal placement of samples should occur at the points where the pdf is steepest. This contradiction is resolved when one realizes that only the original samples are independent. The pdf samples are obtained by taking differences of adjacent PF values, so neighboring samples are anticorrelated. Thus, one cannot claim that for estimating the PSE, the optimal placement of trials is at the steep portion of the pdf.

Wichmann and Hill's Biased Goodness-of-Fit

Whenever I fit a psychometric function to data I also calculate the goodness-of-fit. The second half of Wichmann and Hill (2001a) is devoted to that topic and I have some comments on their findings. There are two standard methods for calculating the goodness-of-fit (Press, Teukolsky, Vetterling, & Flannery, 1992). The first method is to calculate the X^2 statistic, as given by

$$X^2 = \sum_i \frac{\left[p\left(\text{data}_i\right) - P_i\right]^2}{\text{var}\left(P_i\right)}, \tag{46}$$

where $p(\text{data}_i)$ and P_i are the percent correct of the datum and the PF at level i. The binomial $\text{var}(P_i)$ is our old friend:

$$\text{var}\left(P_i\right) = \frac{P_i\left(1 - P_i\right)}{N_i}, \tag{47}$$

where N_i is the number of trials at level i. The X^2 is of interest because the binomial distribution is asymptotically a Gaussian, so that the X^2 statistic asymptotically has a chi-square distribution. Alternatively, one can write X^2 as

$$X^2 = \sum_i \text{residual}_i^2 \tag{48}$$

where the residuals are the difference between the predicted and observed probabilities divided by the standard error of the predicted probabilities, $SE_i = [\text{var}(P_i)]^{0.5}$:

$$\text{residual}_i = \frac{P\left(\text{data}_i\right) - P_i}{SE_i}. \tag{49}$$

The second method is to use the normalized log-likelihood that Wichmann and Hill call the deviance, D. I will use the letters LL, to remind the reader that I am dealing with log likelihood:

$$LL = 2\sum_i N_i\, p(\text{data}_i)\ln\frac{p(\text{data}_i)}{P_i} + N_i\left[1 - p(\text{data}_i)\right]\ln\left[\frac{1 - p(\text{data}_i)}{1 - P_i}\right]. \tag{50}$$

Similar to the X^2 statistic, LL asymptotically also has a chi-square distribution.

In order to calculate the goodness-of-fit from the chi-square statistic, one usually makes the assumption that we are dealing with a linear model for P_i. A linear model means that the various parameters controlling the shape of the psychometric function (like α, β, γ in Equation 1 for the Weibull function) should contribute linearly. However, Equations 1, 8, and 12 show that only γ contributes linearly. Even though the PF function is not linear in its parameters, one typically makes the linearity assumption anyway, in order to use the chi-square goodness-of-fit tables. The purpose of this section is to examine the validity of the linearity assumption.

The chi-square goodness-of-fit distribution for a linear model is tabulated in most statistics books. It is given by the Gamma function (Press et al., 1992):

$$\text{prob\_chisq} = \text{Gamma}\left(df/2,\ \chi^2/2\right) \tag{51}$$

where the degrees of freedom, df, is the number of data points minus the number of parameters. The Gamma function is a standard function (called Gammainc in Matlab). In order to check that the Gamma function is functioning properly, I often check that for a reasonably large number of degrees of freedom (like $df > 10$), the chi-square distribution is close to Gaussian with a mean $\approx df$ and the standard deviation $\approx (2df)^{0.5}$. As an example for $df = 18$, we have Gamma(9, 9) = 0.54, which is very close to the asymptotic value of 0.5. To check the standard deviation, we have Gamma [18/2, (18 + 6)/2] = .845, which is indeed close to the value of 0.841 expected for a unity z-score.

With a sparse number of trials, or with probabilities near unity or with nonlinear models (all of which occur in fitting psychometric functions), one might worry about the validity of using the chi-square distribution (from standard tables or Equation 51). Monte Carlo simulations would be more trustworthy. That is what Wichmann and Hill (2001a) use for the log likelihood deviance, LL. In Monte Carlo simulations, one replicates the experimental conditions and uses different randomly generated data sets based on binomial statistics. Wichmann and Hill do 10,000 runs for each condition. They calculate LL for each run and then plot a histogram of the Monte Carlo simulations for comparison with the expected chi-square given by Equation 50, based on the chi-square distribution.

Their most surprising finding is the strong bias displayed in their Figures 7 and 8. Their solid line is the chi-square distribution from Equation 51. Both Figures 7a and 8a are for a total of 120 trials per run, with 60 levels and 2 trials for each level. In Figure 7a, with a strong upward bias, the trials span the probability interval [0.52 , 0.85]; in Figure 8a, with a strong downward bias, the interval is [0.72, 0.99]. That relatively innocuous shift from testing at lower probabilities to testing at higher probabilities made a dramatic shift in the bias of the LL distribution relative to the chi-square distribution. The shift perplexed me, so I decided to investigate both X^2 and likelihood for different numbers of trials and different probabilities for each level.

Matlab Code 2 in the Appendix is the program that calculates LL and X^2 for a PF ranging from P = 50% to 100%, and for 1, 2, 4, 8, or 16 trials at each level. The Matlab code calculates Equations 46 and 50 by enumerating all possible outcomes, rather than by doing Monte Carlo simulations. Consider the contribution to chi square from one level in the sum of Equation 46:

$$X_i^2 = \frac{\left[p(\text{data}_i) - P_i\right]^2}{\dfrac{P_i(1 - P_i)}{N_i}} = \frac{\left(O_i - E_i\right)^2}{E_i\left(1 - P_i\right)}, \tag{52}$$

where $O_i = N_i\, p(\text{data}_i)$ and $E_i = N_i P_i$. The mean value of X_i^2 is obtained by doing a weighted sum over all possible outcomes for O_i. The weighting, wt_i, is the expected probability from binomial statistics:

$$wt_i = \frac{N_i!}{O_i!\left(N_i - O_i\right)!} \cdot P_i^{O_i}\left(1 - P_i\right)^{N_i - O_i}. \tag{53}$$

The expected value $<X_i^2>$ is

$$< X_i^2 > = \sum\nolimits_{O_i} wt_i \frac{\left(O_i - E_i\right)^2}{E_i\left(1 - P_i\right)}. \tag{54}$$

A bit of algebra based on $\sum_{O_i} wt_i = 1$ gives a surprising result:

$$< X_i^2 > = 1. \tag{55}$$

That is, each term of the X^2 sum has an expectation value of unity, independent of N_i and independent of P! Having each deviant contribute unity to the sum in Equation 46 is desired, since $E_i = N_i P_i$ is the exact value rather than a fitted value based on the sampled data. In Figure 4, the horizontal line is the contribution to chi square for any probability level. I had wrongly thought that one needed N_i to be fairly big before each term contributed a unity amount, on the average, to X^2. It happens with any N_i.

If we try the same calculation for each term of the summation in Equation 50 for LL, we have

$$< \mathrm{LL}_i > = 2\sum_{O_i} wt_{O_i}\left[O_i \ln\left(\frac{O_i}{E_i}\right) + \left(N_i - O_i\right)\ln\left(\frac{N_i - O_i}{N_i - E_i}\right)\right]. \tag{56}$$

Unfortunately, there is no simple way to do the summations as was done for X^2, so I resort to the program of Matlab Code 2. The output of the program is shown in Figure 4. The five curves are for $N_i = 1, 2, 4, 8$, and 16 as indicated. It can be seen that for low values of P_i, near .5, the contribution of each term is greater than 1, and that for high values (near 1), the contribution is less than 1. As N_i increases, the contribution of most levels gets closer to 1, as expected. There are, however, disturbing deviations from unity at high levels of P_i. An examination of the case $N_i = 2$ is useful, since it is the case for which Wichmann and Hill (2001a) showed strong biases in their Figures 7 and 8 (left-hand panels). This examination will show why Figure 4 has the shape shown.

I first examine the case where the levels are biased toward the lower asymptote. For $N_i = 2$ and $P_i = 0.5$, the weights in Equation 53 are 1/4, 1/2, and 1/4 for $O_i = 0$, 1, or 2 and $E_i = N_i P_i = 1$. Equation 56 simplifies, since the factors with $O_i = 0$ and $O_i = 1$ vanish from the first term and the factors with $O_i = 2$ and $O_i = 1$ vanish from the second term. The $O_i = 1$ terms vanish because log(1/1) = 0. Equation 56 becomes

$$\begin{aligned} < \mathrm{LL}_i > &= 2 * 0.25\left[2\ln(2) + 2 * 2\ln(2)\right] \\ &= 2\ln(2) = 1.386. \end{aligned} \tag{57}$$

Figure 4 shows that this is precisely the contribution of each term at $P_i = .5$. Thus, if the PF levels were skewed to the low side, as in Wichmann and Hill's Figure 7 (left panel), the present analysis predicts that the LL statistic would be biased to the high side. For the 60 levels in Figure 7 (left panel), their deviance statistic was biased about 20% too high, which is compatible with our calculation.

Now I consider the case where levels are biased near the upper asymptote. For $P_i = 1$ and any value of N_i, the weights are zero because of the $1-P_i$ factor in Equation 53 except for $O_i = N_i$ and $E_i = N_i$. Equation 56 vanishes either because of the weight or because of the log term, in agreement with Figure 4. Figure 4 shows that for levels of $P_i > .85$, the contribution to chi-square is less than unity. Thus if the PF levels were skewed to the high side, as in Wichmann and Hill's Figure 8 (left panel), we would expect the LL statistic to be biased below unity, as they found.

I have been wondering about the relative merits of X^2 and log likelihood for many years. I have always appreciated the simplicity and intuitive nature of X^2, but statisticians seem to prefer likelihood. I do not doubt the arguments of King-Smith et al. (1994) and Kontsevich and Tyler (1999), who advocate using mean likelihood in a

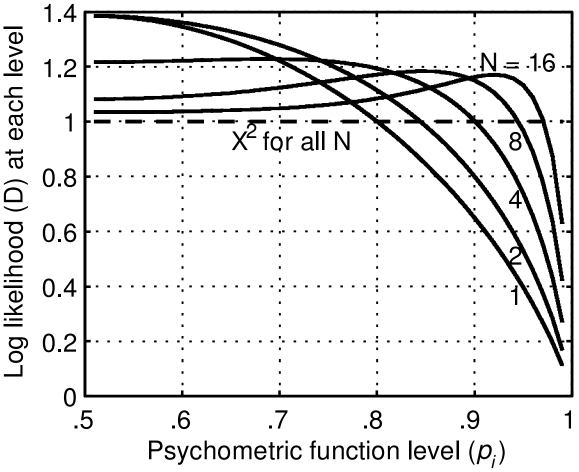

Figure 4. The bias in goodness-of-fit for each level of 2AFC data. The abscissa is the probability correct at each level P_i. The ordinate is the contribution to the goodness-of-fit metric (X^2 or log likelihood deviance). In linear regression with Gaussian noise, each term of the chi-square summation is expected to give a unity contribution if the true rather than sampled parameters are used in the fit, so that the expected value of chi square equals the number of data points. With binomial noise, the horizontal dashed line indicates that the X^2 metric (Equation 52) contributes exactly unity, independent of the probability and independent of the number of trials at each level. This contribution was calculated by a weighted sum (Equations 53–54) over all possible outcomes of data, rather than by doing Monte Carlo simulations. The program that generated the data is included in the Appendix as Matlab Code 2. The contribution to log likelihood deviance (specified by Equation 56) is shown by the five curves labeled 1–16. Each curve is for a specific number of trials at each level. For example, for the curve labeled 2 with 2 trials at each level tested, the contribution to deviance was greater than unity for probability levels less than about 85% correct and was less than unity for higher probabilities. This finding explains the goodness-of-fit bias found by Wichmann and Hill (2001a, Figures 7a and 8a).

Bayesian framework for adaptive methods in order to measure the PF. However, for goodness-of-fit, it is not clear that the log likelihood method is better than X^2. The common argument in favor of likelihood is that it makes sense even if there is only one trial at each level. However, my analysis shows that the X^2 statistic is less biased than log likelihood when there is a low number of trials per level. This surprised me. Wichmann and Hill (2001a) offer two more arguments in favor of log likelihood over X^2. First, they note that the maximum likelihood parameters will not minimize X^2. This is not a bothersome objection, since if one does a goodness-of-fit with X^2, one would also do the parameter search by minimizing X^2 rather than maximizing likelihood. Second, Wichmann and Hill claim that likelihood, but not X^2, can be used to assess the significance of added parameters in embedded models. I doubt that claim since it is standard to use χ^2 for embedded models (Press et al., 1992), and I cannot see that X^2 should be any different.

Finally, I will mention why goodness-of-fit considerations are important. First, if one consistently finds that one's data have a poorer goodness-of-fit than do simulations, one should carefully examine the shape of the PF fitting function and carefully examine the experimental methodology for stability of the subjects' responses. Second, goodness-of-fit can have a role in weighting multiple measurements of the same threshold. If one has a reliable estimate of threshold variance, multiple measurements can be averaged by using an inverse variance weighting. There are three ways to calculate threshold variance: (1) One can use methods that depend only on the best-fitting PF (see the discussion of inverse of Hessian matrix in Press et al., 1992). The asymptotic formula such as that derived in Equations 39–43 provides an example for when only threshold is a free parameter. (2) One can multiply the variance of method (1) by the reduced chi-square (chi-square divided by the degrees of freedom) if the reduced chi-square is greater than one (Bevington, 1969; Klein, 1992). (3) Probably the best approach is to use bootstrap estimates of threshold variance based on the data from each run (Foster & Bischof, 1991; Press et al., 1992). The bootstrap estimator takes the goodness-of-fit into account in estimating threshold variance. It would be useful to have more research on how well the threshold variance of a given run correlates with threshold accuracy (goodness-of-fit) of that run. It would be nice if outliers were strongly correlated with high threshold variance. I would not be surprised if further research showed that because the fitting involves nonlinear regression, the optimal weighting might be different from simply using the inverse variance as the weighting function.

Wichmann and Hill, and Strasburger: Lapses and Bias Calculation

The first half of Wichmann and Hill (2001a) is concerned with the effect of lapses on the slope of the PF. "Lapses," also called "finger errors," are errors to highly visible stimuli. This topic is important, because it is typically ignored when one is fitting PFs. Wichmann and Hill (2001a) show that improper treatment of lapses in parametric PF fitting can cause sizeable errors in the values of estimated parameters. Wichmann and Hill (2001a) show that these considerations are especially important for estimation of the PF slope. Slope estimation is central to Strasburger's (2001b) article on letter discrimination. Strasburger's (2001b) Figures 4, 5, 9, and 10 show that there are substantial lapses even at high stimulus levels. Strasburger's (2001b) maximum likelihood fitting program used a non-zero lapse parameter of $\lambda = 0.01\%$ (H. Strasburger, personal communication, September 17, 2001). I was worried that this value for the lapse parameter was too small to avoid the slope bias, so I carried out a number of simulations similar to those of Wichmann and Hill (2001a), but with a choice of parameters relevant to Strasburger's situation. Table 2 presents the results of this study.

Since Strasburger used a 10AFC task with the PF ranging from 10% to 100% correct, I decided to use discrimination PFs going from 0% to 100% rather than the 2AFC PF of Wichmann and Hill (2001a). For simplicity I decided to have just three levels with 50 trials per level, the same as Wichmann and Hill's example in their Figure 1. The PF that I used had 25, 42, and 50 correct responses at levels placed at $x = 0$, 1, and 6 (columns labeled "No Lapse"). A lapse was produced by having 49 instead of 50 correct at the high level (columns labeled "Lapse"). The data were fit in six different ways: Three PF shapes were used, probit (cumulative normal), logit, and Weibull corresponding to the rows of Table 2; and two error metrics were used, chi-square minimization and likelihood

Table 2
The Effect of Lapses on the Slope of Three Psychometric Functions

Psychometric	$\lambda = .01$ No Lapse		$\lambda = .0001$ No Lapse		$\lambda = .01$ Lapse		$\lambda = .0001$ Lapse	
Function	L	χ^2	L	χ^2	L	χ^2	L	χ^2
Probit	1.05	1.05	1.00	0.99	1.05	1.05	0.36	0.34
Logit	1.76	1.76	1.66	1.66	1.76	1.76	0.84	0.72
Weibull	1.01	1.01	0.97	0.97	1.01	1.01	0.26	0.26

Note—The columns are for different lapse rates and for the presence or absence of lapses. The 12 pairs of entries in the cells are the slopes; the left and right values correspond to likelihood maximization (L) and χ^2 minimization, respectively.

maximization, corresponding to the paired entries in the table. In addition, two lapse parameters were used in the fit: $\lambda = 0.01$ (first and third pairs of data columns) and $\lambda = 0.0001$ (second and fourth pairs of columns). For this discrimination task, the lapses were made symmetric by setting $\gamma = \lambda$ in Equation 1A so that the PF would be symmetric around the midpoint.

The Matlab program that produced Table 2 is included as Matlab Code 3 in the Appendix. The details on how the fitting was done and the precise definitions of the fitting functions are given in the Matlab code. The functions being fit are

$$\text{probit} \quad P = .5 + .5\,\mathrm{erf}\left(\frac{z}{\sqrt{2}}\right) \quad \text{(Equation 3B)} \quad (58\text{A})$$

$$\text{logit} \quad P = \frac{1}{1+\exp(-z)} \quad (58\text{B})$$

$$\text{Weibull} \quad P = 1-\exp\left[-\exp(z)\right] \quad \text{(Equation 13)} \quad (58\text{C})$$

with $z = p_2(y - p_1)$. The parameters p_1 and p_2, representing the threshold and slope of the PF, are the two free parameters in the search. The resulting slope parameters are reported in Table 2. The left entry of each pair is the result of the likelihood maximization fit, and the right entry is that of the chi-square minimization. The results showed that (1) When there were no lapses (50 out of 50 correct at the high level), the slope did not strongly depend on the lapse parameter. (2) When the lapse parameter was set to $\lambda = 1\%$, the PF slope did not change when there was a lapse (49 out of 50). (3) When $\lambda = 0.01\%$, the PF slope was reduced dramatically when a single lapse was present. (4) For all cases except the fourth pair of columns, there was no difference in the slope estimate between maximizing likelihood or minimizing chi-square as would be expected, since in these cases the PF fit the data quite well (low chi-square). In the case of a lapse with $\lambda = 0.01\%$ (fourth pair of columns), chi-square is large (not shown), indicating a poor fit, and there is an indication in the logit row that chi-square is more sensitive to the outlier than is the likelihood function. In general, chi-square minimization is more sensitive to outliers than is likelihood maximization. Our results are compatible with those of Wichmann and Hill (2001a). Given this discussion, one might worry that Strasburger's estimated slopes would have a downward bias, since he used $\lambda = 0.0001$ and he had substantial lapses. However, his slopes were quite high. It is surprising that the effect of lapses did not produce lower slopes.

In the process of doing the parameter searches that went into Table 2, I encountered the problem of "local minima," which often occurs in nonlinear regression but is not always discussed. The PFs depend nonlinearly on the parameters, so both chi-square minimization and likelihood maximization can be troubled by this problem whereby the best-fitting parameters at the end of a search depend on the starting point of the search. When the fit is good (a low chi-square), as is true in the first three pairs of data columns of Table 2, the search was robust and relatively insensitive to the initial choice of parameter. However, in the last pair of columns, the fit was poor (large chi-square) because there was a lapse but the lapse parameter was too small. In this case, the fit was quite sensitive to choice of initial parameters, so a range of initial values had to be explored to find the global minimum. As can be seen in the Matlab Code 3 in the Appendix, I set the initial choice of slope to be 0.3, since an initial guess in that region gave the overall lowest value of chi-square.

Wichmann and Hill's (2001a) analysis and the discussion in this section raise the question of how one decides on the lapse rate. For an experiment with a large number of trials (many hundreds), with many trials at high stimulus strengths, Wichmann and Hill (2001a, Figure 5) show that the lapse parameter should be allowed to vary while one uses a standard fitting procedure. However, it is rare to have sufficient data to be able to estimate slope and lapse parameters as well as threshold. One good approach is that of Treutwein and Strasburger (1999) and Wichmann and Hill (2001a), who use Bayesian priors to limit the range of λ. A second approach is to weight the data with a combination of binomial errors based on the observed as well as the expected data (Klein, 2002) and fix the lapse parameter to zero or a small number. Normally the weighting is based solely on the expected analytic PF rather than the data. Adding in some variance due to the observed data permits the data with lapses to receive lesser weighting. A third approach, proposed by Manny and Klein (1985), is relevant to testing infants and clinical populations, in both of which cases the lapse rate might be high, with the stimulus levels well separated. The goal in these clinical studies is to place bounds on threshold estimates. Manny and Klein used a step-like PF with the assumption that the slope was steep enough that only one datum lay on the sloping part of the function. In the fitting procedure, the lapse rate was given by the average of the data above the step. A maximum likelihood procedure with threshold as the only free parameter (the lapse rate was constrained as just mentioned) was able to place statistical bounds on the threshold estimates.

Biased Slope in Adaptive Methods

In the previous section, I examined how lapses produce a downward slope bias. Now I will examine factors that can produce an upward slope bias. Leek, Hanna, and Marshall (1992) carried out a large number of 2AFC, 3AFC, and 4AFC simulations of a variety of Brownian staircases. The PF that they used to generate the data and to fit the data was the power law d' function ($d' = x_t^b$). (I was pleased that they called the d' function the "psychometric function.") They found that the estimated slope, b, of the PF was biased high. The bias was larger for runs with fewer trials and for PFs with shallower slopes or more closely spaced levels. Two of the papers in this special issue were centrally involved with this topic. Strasburger (2001b) used a 10AFC task for letter discrimination and found slopes that were more than double the slopes in previous studies. I worried that the high slopes might be due to a bias caused by the methodology. Kaernbach (2001b) provides a mechanism that could explain the slope bias found in adaptive procedures. Here I will summarize my thoughts on this topic. My motivation goes well beyond the particular issue of a slope bias associated with adaptive methods. I see it as an excellent case study that provides many lessons on how subtle, seemingly innocuous methods can produce unwanted biases.

Strasburger's steep slopes for character recognition. Strasburger (2001b) used a maximum likelihood adaptive procedure with about 30 trials per run. The task was letter discrimination with 10 possible peripherally viewed letters. Letter contrast was varied on the basis of a likelihood method, using a Weibull function with $\beta = 3.5$ to obtain the next test contrast and to obtain the threshold at the end of each run. In order to estimate the slope, the data were shifted on a log axis so that thresholds were aligned. The raw data were then pooled so that a large number of trials were present at a number of levels. Note that this procedure produces an extreme concentration of trials at threshold. The bottom panels of Strasburger's Figures 4 and 5 show that there are less than half the number of trials in the two bins adjacent to the central bin, with a bin separation of only 0.01 log units (a 2.3% contrast change). A Weibull function with threshold and slope as free parameters was fit to the pooled data, using a lower asymptote of $\gamma = 10\%$ (because of the 10AFC task) and a lapse rate of $\lambda = 0.01\%$ (a 99.990% correct upper asymptote). The PF slope was found to be $\beta = 5.5$. Since this value of β is about twice that found regularly, Strasburger's finding implies at least a four-fold variance reduction. The asymptotic (large N) variance for the 10AFC task is $1.75/(N\beta^2) = 0.058/N$ (Klein, 2001). For a run of 25 trials, the variance is var $= 0.058/25 = 0.0023$. The standard error is $SE = \text{sqrt(var)} \approx .05$. This means that in just 25 trials, one can estimate threshold with a 5% standard error, a low value. That low value is sufficiently remarkable that one should look for possible biases in the procedure.

Before considering a multiplicity of factors contributing to a bias, it should be noted that the large values of β could be real for seeing the tiny stimuli used by Strasburger (2001b). With small stimuli and peripheral viewing, there is much spatial uncertainty, and uncertainty is known to elevate slopes. A question remains, of whether the uncertainty is sufficient to account for the data.

There are many possible causes of the upward slope bias. My intuition was that the early step of shifting the PF to align thresholds was an important contributor to the bias. As an example of how it would work, consider the five-point PF with equal level spacing used by Miller and Ulrich (2001): $P(i)$ = 1%, 12.22%, 50%, 87.78%, and 99%. Suppose that because of binomial variability, the middle point was 70% rather than 50%. Then the threshold would be shifted to between Levels 2 and 3, and the slope would be steepened. Suppose, on the other hand, that the variability causes the middle point to be 30% rather than 50%. Now the threshold would be shifted to between Levels 3 and 4, and the slope would again be steepened. So any variability in the middle level causes steepening. Variability at the other levels has less effect on slope. I will now compare this and other candidates for slope bias.

Kaernbach's explanation of the slope bias. Kaernbach (2001b) examines staircases based on a cumulative normal PF that goes from 0% to 100%. The staircase rule is Brownian, with one step up or down for each incorrect or correct answer, respectively. Parametric and nonparametric methods were used to analyze the data. Here I will focus on the nonparametric methods used to generate Kaernbach's Figures 2–4. The analysis involves three steps.

First, the data are monotonized. Kaernbach gives two reasons for this procedure: (1) Since the true psychometric function is expected to be monotonic, it seems appropriate to monotonize the data. This also helps remove noise from nonparametric threshold or slope estimates. (2) "Second, and more importantly, the monotonicity constraint improves the estimates at the borders of the tested signal range where only few tests occur and admits to estimate the values of the PF at those signal levels that have not been tested (i.e., above or below the tested range). For the present approach this extrapolation is necessary since the averaging of the PF values can only be performed if all runs yield PF estimates for all signal levels in question" (Kaernbach, 2001b, p. 1390).

Second, the data are extrapolated, with the help of the monotonizing step as mentioned in the quote of the preceding paragraph. Kaernbach (2001b) believes it is essential to extrapolate. However, as I shall show, it is possible to do the simulations without the extrapolation step. I will argue in connection with my simulations that the extrapolation step may be the most critical step in Kaernbach's method for producing the bias he finds.

Third, a PF is calculated for each run. The PFs are then averaged across runs. This is different from Strasburger's method, in which the data are shifted and then, rather than average the PFs of each run, the total number correct and incorrect at each level are pooled. Finally, the PF is generated on the basis of the pooled data.

Simulations to clarify contributions to the slope bias. A number of factors in Kaernbach's and Strasburger's procedures could contribute to biased slope estimate. In Strasburger's case, there is the shifting step. One might think this special to Strasburger, but in fact it occurs in many methods, both parametric and nonparametric. In parametric methods, both slope and threshold are typically estimated. A threshold estimate that is shifted away from its true value corresponds to Strasburger's shift. Similarly, in the nonparametric method of Miller and Ulrich (2001), the distribution is implicitly allowed to shift in the process of estimating slope. When Kaernbach (2001b) implemented Miller and Ulrich's method, he found a large upward slope bias. Strasburger's shift step was more explicit than that of the other methods, in which the shift is implicit. Strasburger's method allowed the resulting slope to be visualized as in his figures of the raw data after the shift.

I investigated several possible contributions to the slope bias: (1) shifting, (2) monotonizing, (3) extrapolating, and (4) averaging the PF. Rather than simulations, I did enumerations, in which all possible staircases were analyzed with each staircase consisting of 10 trials, all starting at the same point. To simplify the analysis and to reduce the number of assumptions, I chose a PF that was flat (slope of zero) at $P = 50\%$. This corresponds to the case of a normal PF, but with a very small step size. The Matlab program that does all the calculations and all the figures is included as Matlab Code 4. There are four parts to the code: (1) the main program, (2) the script for monotonizing, (3) the script for extrapolating, (4) the script for either averaging the PFs or totaling up the raw responses. The Matlab code in the Appendix shows that it is quite easy to find a method for averaging PFs even when there are levels that are not tested. Since the annotated programs are included, I will skip the details here and just offer a few comments.

The output of the Matlab program is shown in Figure 5. The left and right columns of panels show the PF for the case of no shifting and with shifting, respectively. The shifting is accomplished by estimating threshold as the mean of all the levels, a standard nonparametric method for estimating threshold. That mean is used as the amount by which the levels are shifted before the data from all runs are pooled. The top pair of panels is for the case of averaging the raw data; the middle panels show the results of averaging following the monotonizing procedure. The monot-

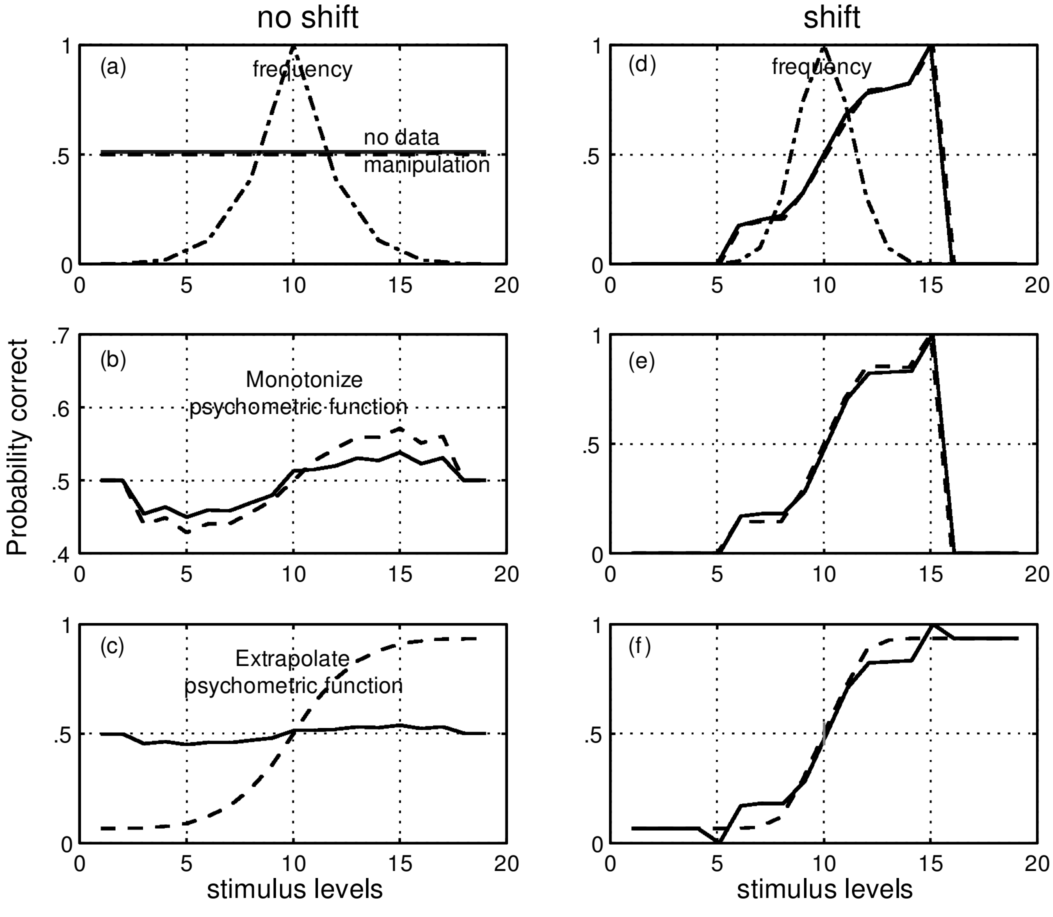

Figure 5. Slope bias for staircase data. Psychometric functions are shown following four types of processing steps: shifting, monotonizing, extrapolating, and type of averaging. In all panels, the solid line is for averaging the raw data of multiple runs before calculating the PF. The dashed line is for first calculating the PF and then averaging the PF probabilities. Panel a shows the initial PF is unchanged if there is no shifting, maximizing, or extrapolating. For simplicity, a flat PF, fixed at $P = 50\%$, was chosen. The dot–dashed curve is a histogram showing the percentage of trials at each stimulus level. Panel b shows the effect of monotonizing the data. Note that this one panel has a reduced ordinate to better show the details of the data. Panel c shows the effect of extrapolating the data to untested levels. Panels d, e, and f are for the cases a, b, and c where in addition a shift operation is done to align thresholds before the data are averaged across runs. The results show that the shift operation is the most effective analysis step for producing a bias. The combination of monotonizing, extrapolating, and averaging PFs (panel c) also produces a strong upward bias of slope.

onizing routine took some thought, and I hope its presence in the Appendix (Matlab Code 4) will save others much trouble. The lower pair of panels shows the results of averaging after both monotonizing and extrapolating.

In each panel, the solid curve shows the average PF obtained by combining the correct trials and the total trials across runs for each level and taking their ratio to get the PF. The dashed curve shows the average PF resulting from averaging the PFs of each run. The latter method is the more important one, since it simulates single short runs. In the upper pair of panels, the dashed and solid curves are so close to each other that they look like a single curve. In the upper pair, I show an additional dot–dashed line, with right–left symmetry, that represents the total number of trials. The results in the plots are as follows.

Placement of trials. The effect of the shift on the number of trials at each level can be seen by comparing the dot–dashed lines in the top two panels. The shift narrows the distribution significantly. There are close to zero trials more than three steps from threshold in panel d with the shift, even though the full range is ±10 steps. The curve showing the total number of trials would be even sharper if I had used a PF with positive slope rather than a PF that is constant at 50%.

Slope bias with no monotonizing. The top left panel shows that with no shift and no monotonizing, there is no slope bias. The PF is flat at 50%. The introduction of a shift produces a dramatic slope bias, going from PF = 20% to 80% as the stimulus goes from Levels 8 to 12. There was no dependence on the type of averaging.

Effect of monotonizing on slope bias. Monotonizing alone (middle panel on left) produced a small slope bias that was larger with PF averaging than with data averaging. Note that the ordinate has been expanded in this one case, in order to better illustrate the extent of the bias. The extreme amount of steepening (right panel) that is produced by shifting was not matched by monotonizing.

Effect of extrapolating on slope bias. The lower left panel shows the dramatic result that the combination of all three of Kaernbach's methods—monotonizing, extrapolating, and averaging PFs—does produce a strong slope bias for these Brownian staircases. If one uses Strasburger's type of averaging, in which the data are pooled before the PF is calculated, then the slope bias is minimal. This type of averaging is equivalent to a large increase in the number of trials.

These simulations show that it is easy to get an upward slope bias from Brownian data with trials concentrated hear one point. An important factor underlying this bias is the strong correlation in staircase levels, discussed by Kaernbach (2001b). A factor that magnifies the slope bias is that when trials are concentrated at one point, the slope estimate has a large standard error, allowing it to shift easily. Our simulations show that one can produce biased slope estimates either by a procedure (possibly implicit) that shifts the threshold, or by a procedure that extrapolates data to untested levels. This topic is of more academic than practical interest, since anyone interested in the PF slope should use an adaptive algorithm like the Ψ method of Kontsevich and Tyler (1999), in which trials are placed at separated levels for estimating slope. The slope bias that is produced by the seemingly innocuous extrapolation or shifting step is a wonderful reminder of the care that must be taken when one employs psychometric methodologies.

SUMMARY

Many topics have been in this paper, so a summary should serve as a useful reminder of several highlights. Items that are surprising, novel, or important are italicized.

I. Types of PFs

1. There are two methods for compensating for the PF lower asymptote, also called the *correction for bias*:

1.1 Do the compensation in probability space. Equation 2 specifies $p(x) = [P(x)-P(0)] / [1-P(0)]$, where $P(0)$, the lower asymptote of the PF, is often designated by the parameter γ. With this method, threshold estimates vary as the lower asymptote varies (a failure of the high-threshold assumption).

1.2 *Do the compensation in* z*-score space for yes/no tasks. Equation 5 specifies* $d'(x) = z(x) - z(0)$. *With this method, the response measure* d′ *is identical to the metric used in signal detection theory*. This way of looking at psychometric functions is not familiar to many researchers.

2. 2AFC is not immune to bias. It is not uncommon for the observer to be biased toward Interval 1 or 2 when the stimulus is weak. Whereas the criterion bias for a yes/no task [$z(0)$ in Equation 5] affects d' linearly, the interval bias in 2AFC affects d' quadratically and therefore it has been thought small. *Using an example from Green and Swets (1966), I have shown that this 2AFC bias can have a substantial effect on* d' and on threshold estimates, a different conclusion from that of Green and Swets. *The interval bias can be eliminated by replotting the 2AFC discrimination PF, using the percent correct Interval 2 judgments on the ordinate and the Interval 2 minus Interval 1 signal strength on the abscissa.* This 2AFC PF goes from 0% to 100%, rather than from 50% to 100%.

3. Three distinctions can be made regarding PFs: yes/no versus forced choice, detection versus discrimination, adaptive versus constant stimuli. In all of the experimental papers in this special issue, the use of forced choice adaptive methods reflects their widespread prevalence.

4. Why is 2AFC so popular given its shortcomings? A common response is that 2AFC minimizes bias (but note Item 2 above). A less appreciated reason is that *many adaptive methods are available for forced choice methods, but that objective yes/no adaptive methods are not common.* By "objective," I mean a signal detection method with sufficient blank trials to fix the criterion. Kaernbach (1990) proposed an objective yes/no, up–down staircase with rules as simple as these: Randomly intermix blanks and

signal trials, decrease the level of the signal by one step for every correct answer (hits or correct rejections), and increase the level by three steps for every wrong answer (false alarms or misses). The minimum d' is obtained when the numbers of "yes" and "no" responses are about equal (ROC negative diagonal). *A bias of imbalanced "yes" and "no" responses is similar to the 2AFC interval bias discussed in Item 2.* The appropriate balance can be achieved by giving bias feedback to the observer.

5. *Threshold should be defined on the basis of a fixed d' level rather than the 50% point of the PF.*

6. *The connection between PF slope on a natural logarithm abscissa and slope on a linear abscissa is as follows: $slope_{lin} = slope_{log}/x_t$ (Equation 11), where x_t is the stimulus strength in threshold units.*

7. Strasburger's (2001a) maximum PF slope has the advantage that it relates PF slope using a probability ordinate, β', to the low stimulus strength log–log slope of the d' function, β. It has the further advantage of relating the slopes of a wide variety of PFs: Weibull, logistic, Quick, cumulative normal, hyperbolic tangent, and signal detection d'. The main difficulty with maximum slope is that quite often, threshold is defined at a point different from the maximum slope point. Typically, the point of maximum slope is lower than the level that gives minimum threshold variance.

8. *A surprisingly accurate connection was made between the 2AFC Weibull function and the Stromeyer–Foley d' function* given in Equation 20. For all values of the Weibull β parameter and for all points on the PF, the maximum difference between the two functions is .0004 on a probability ordinate going from .50 to 1.00. For this good fit, *the connection between the d' exponent b and the Weibull exponent β is $b/\beta = 1.06$.* This ratio differs from $b/\beta \approx 0.8$ (Pelli, 1985) and $b/\beta = 0.88$ (Strasburger, 2001a), because our d' function saturates at moderate d' values just as the Weibull saturates and as experimental data saturate.

II. Experimental Methods for Measuring PFs

1. The 2AFC and objective yes/no threshold variances were compared, using signal detection assumptions. *For a fixed total percent correct, the yes/no and the 2AFC methods have identical threshold variances when using a d' function given by $d' = c_t{}^b$.* The optimum variance of $1.64/(Nb^2)$ occurs at $P = 94\%$, where N is the number of trials. If N is the number of stimulus presentations, then, for the 2AFC task, the variance would be doubled to $3.28/(Nb^2)$. *Counting stimulus presentations rather than trials can be important for situations in which each presentation takes a long time, as in the smell and taste experiments of Linschoten et al. (2001).*

2. The equality of the 2AFC and yes/no threshold variance leads to a paradox whereby observers could convert the 2AFC experiment to an objective yes/no experiment by closing their eyes in the first 2AFC interval. Paradoxically, the eyes' shutting would leave the threshold variance unchanged. This paradox was resolved when a more realistic d'PF with saturation was used. In that case, the yes/no variance was slightly larger than the 2AFC variance for the same number of trials.

3. *Several disadvantages of the 2AFC method are that (a) 2AFC has an extra memory load, (b) modeling probability summation and uncertainty is more difficult than for yes/no, (c) multiplicative noise (ROC slope) is difficult to measure in 2AFC, (d) bias issues are present in 2AFC as well as in objective yes/no, and (e) threshold estimation is inefficient in comparison with methods with lower asymptotes that are less than 50%.*

4. *Kaernbach (2001b) suggested that some 2AFC problems could be alleviated by introducing an extra "don't know" response category. A better alternative is to give a high or low confidence response in addition to choosing the interval.* I discussed how the confidence rating would modify the staircase rules.

5. *The efficiency of the objective yes/no method could be increased by increasing the number of response categories (ratings) and increasing the number of stimulus levels (Klein, 2002).*

6. *The simple up–down (Brownian) staircase with threshold estimated by averaging levels was found to have nearly optimal efficiency in agreement with Green (1990). It is better to average all levels (after about four reversals of initial trials are thrown out) rather than average an even number of reversals.* Both of these findings were surprising.

7. As long as the starting point is not too distant from threshold, Brownian staircases with their moderate inertia have advantages over the more prestigious adaptive likelihood methods. *At the beginning of a run, likelihood methods may have too little inertia and can jump to low stimulus strengths before the observer is fully familiar with the test target. At the end of the run, likelihood methods may have too much inertia and resist changing levels even though a nonstationary threshold has shifted.*

8. The PF slope is useful for several reasons. It is needed for estimating threshold variance, for distinguishing between multiple vision models, and for improved estimation of threshold. I have discussed PF slope as well as threshold. Adaptive methods are available for measuring slope as well as threshold.

9. *Improved goodness-of-fit rules are needed as a basis for throwing out bad runs and as a means for improving estimates of threshold variance.* The goodness-of-fit metric should look not only at the PF, but also at the sequential history of the run so that mid-run shifts in threshold can be determined. The best way to estimate threshold variance on the basis of a single run of data is to carry out bootstrap calculations (Foster & Bischof, 1991; Wichmann & Hill, 2001b). Further research is needed in order to determine the optimal method for weighting multiple estimates of the same threshold.

10. *The method should depend on the situation.*

III. Mathematical Methods for Analyzing PFs

1. Miller and Ulrich's (2001) *nonparametric Spearman–Kärber (SK) analysis can be as good as and often better

than a parametric analysis. An important *limitation of the SK analysis is that it does not include a weighting factor that would emphasize levels with more trials*. This would be relevant for staircase methods in which the number of trials was unequally distributed. It would also cause problems with 2AFC detection, because of the asymmetry of the upper and lower asymptote. It need not be a problem for 2AFC discrimination, because, as I have shown, this task can be analyzed better with a negative-going abscissa that allows the ordinate to go from 0% to 100%. *Equations 39–43 provide an analytic method for calculating the optimal variance of threshold estimates for the method of constant stimuli. The analysis shows that the SK method had an optimal variance.*

2. Wichmann and Hill (2001a) carry out a large number of simulations of the chi-square and likelihood goodness-of-fit for 2AFC experiments. They found trial placements where their simulations were upwardly skewed, in comparison with the chi-square distribution based on linear regression. They found other trial placements where their chi-square simulations were downwardly skewed. These puzzling results rekindled my longstanding interest in wanting to know when to trust the chi-square distribution in nonlinear situations and prompted me to *analyze rather than simulate* the biases shown by Wichmann & Hill. To my surprise *I found that the X^2 distribution (Equation 52) had zero bias at any testing level,* even for 1 or 2 trials per level. *The likelihood function (Equation 56), on the other hand, had a strong bias when the number of trials per level was low (Figure 4). The bias as a function of test level was precisely of the form that is able to account for the bias found by Wichmann and Hill (2001a).*

3. Wichmann and Hill (2001a) present a detailed investigation of the strong effect of lapses on estimates of threshold and slope. This issue is relevant to the data of Strasburger (2001b) because of the lapses shown in his data and because a very low lapse parameter ($\lambda = 0.0001$) was used in fitting the data. In my simulations, using PF parameters somewhat similar to Strasburger's situation, I found that the effect of lapses would be expected to be strong, so that Strasburger's estimated slopes should be lower than the actual slopes. However, Strasburger's slopes were high, so the mystery remains of why the lapses did not have a stronger effect. My simulations show that one possibility is *the local minimum problem whereby the slope estimate in the presence of lapses is sensitive to the initial starting guesses for slope in the search routine*. In order to find the global minimum, one needs to use a range of initial slope guesses.

4. One of the most intriguing findings in this special issue was the quite high PF slopes for letter discrimination found by Strasburger (2001b). The slopes might be high because of stimulus uncertainty in identifying small low-contrast peripheral letters. There is also the possibility that these high slopes were due to methodological bias (Leek et al., 1992). Kaernbach (2001b) presents a wonderful analysis of how an upward bias could be caused by trial nonindependence of staircase procedures (not the method Strasburger used). I did a number of simulations to separate out the effects of threshold shift (one of the steps in the Strasburger analysis), PF monotonization, PF extrapolation, and type of averaging (the latter three used by Kaernbach). My results indicated that for experiments such as Strasburger's, the *dominant cause of upward slope bias was the threshold shift. Kaernbach's extrapolation, when coupled with monotonization, can also lead to a strong upward slope bias*. Further experiments and analyses are encouraged, since true steep slopes would be very important in allowing thresholds to be estimated with much fewer trials than is presently assumed. Although none of our analyses makes a conclusive case that Strasburger's estimated slopes are biased, the possibility of a bias suggests that one should be cautious before assuming that the high slopes are real.

5. The slope bias story has two main messages. The obvious one is that *if one wants to measure the PF slope by using adaptive methods, one should use a method that places trials at well separated levels*. The other message has broader implications. *The slope bias shows how subtle, seemingly innocuous methods can produce unwanted effects. It is always a good idea to carry out Monte Carlo simulations of one's experimental procedures, looking for the unexpected.*

REFERENCES

Bevington, P. R. (1969). *Data reduction and error analysis for the physical sciences*. New York: McGraw-Hill.

Carney, T., Tyler, C. W., Watson, A. B., Makous, W., Beutter, B., Chen, C. C., Norcia, A. M., & Klein, S. A. (2000). Modelfest: Year one results and plans for future years. In B. E. Rogowitz, & T. N. Papas (Eds.), *Human vision and electronic imaging V* (Proceedings of SPIE, Vol. 3959, pp. 140-151). Bellingham, WA: SPIE Press.

Emerson, P. L. (1986). Observations on a maximum-likelihood and Bayesian methods of forced-choice sequential threshold estimation. *Perception & Psychophysics*, **39**, 151-153.

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

Foley, J. M. (1994). Human luminance pattern-vision mechanisms: Masking experiments require a new model. *Journal of the Optical Society of America A,* **11**, 1710-1719.

Foster, D. H., & Bischof, W. F. (1991). Thresholds from psychometric functions: Superiority of bootstrap to incremental and probit variance estimators. *Psychological Bulletin,* **109**, 152-159.

Gourevitch, V., & Galanter, E. (1967). A significance test for one parameter isosensitivity functions. *Psychometrika*, **32**, 25-33.

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

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

Hacker, M. J., & Ratcliff, R. (1979). A revised table of d' for M-alternative forced choice. *Perception & Psychophysics*, **26**, 168-170.

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

Hall, J. L. (1983). A procedure for detecting variability of psychophysical thresholds. *Journal of the Acoustical Society of America*, **73**, 663-667.

Kaernbach, C. (1990). A single-interval adjustment-matrix (SIAM) procedure for unbiased adaptive testing. *Journal of the Acoustical Society of America*, **88**, 2645-2655.

Kaernbach, C. (2001a). Adaptive threshold estimation with unforced-choice tasks. *Perception & Psychophysics*, **63**, 1377-1388.

Kaernbach, C. (2001b). Slope bias of psychometric functions derived from adaptive data. *Perception & Psychophysics*, **63**, 1389-1398.

King-Smith, P. E. (1984). Efficient threshold estimates from yes/no procedures using few (about 10) trials. *American Journal of Optometry & Physiological Optics*, **81**, 119.

King-Smith, P. E., Grigsby, S. S., Vingrys, A. J., Benes, S. C., & Supowit, A. (1994). Efficient and unbiased modifications of the QUEST threshold method: Theory, simulations, experimental evaluation and practical implementation. *Vision Research*, **34**, 885-912.

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

Klein, S. A. (1985). Double-judgment psychophysics: Problems and solutions. *Journal of the Optical Society of America*, **2**, 1560-1585.

Klein, S. A. (1992). An Excel macro for transformed and weighted averaging. *Behavior Research Methods, Instruments, & Computers*, **24**, 90-96.

Klein, S. A. (2002). *Measuring the psychometric function*. Manuscript in preparation.

Klein, S. A., & Stromeyer, C. F., III (1980). On inhibition between spatial frequency channels: Adaptation to complex gratings. *Vision Research*, **20** 459-466.

Klein, S. A., Stromeyer, C. F., III, & Ganz, L. (1974). The simultaneous spatial frequency shift: A dissociation between the detection and perception of gratings. *Vision Research*, **15**, 899-910.

Kontsevich, L. L., & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, **39**, 2729-2737.

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics*, **63**, 1279-1292.

Leek, M. R., Hanna, T. E., & Marshall, L. (1991). An interleaved tracking procedure to monitor unstable psychometric functions. *Journal of the Acoustical Society of America*, **90**, 1385-1397.

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

Linschoten, M. R., Harvey, L. O., Jr., Eller, P. M., & Jafek, B. W. (2001). Fast and accurate measurement of taste and smell thresholds using a maximum-likelihood adaptive staircase procedure. *Perception & Psychophysics*, **63**, 1330-1347.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. Cambridge: Cambridge University Press.

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

McKee, S. P., Klein, S. A., & Teller, D. A. (1985). Statistical properties of forced-choice psychometric functions: Implications of probit analysis. *Perception & Psychophysics*, **37**, 286-298.

Miller, J., & Ulrich, R. (2001). On the analysis of psychometric functions: The Spearman–Kärber Method. *Perception & Psychophysics*, **63**, 1399-1420.

Pelli, D. G. (1985). Uncertainty explains many aspects of visual contrast detection and discrimination. *Journal of the Optical Society of America A*, **2**, 1508-1532.

Pelli, D. G. (1987). The ideal psychometric procedures [Abstract]. *Investigative Ophthalmology & Visual Science*, **28** (Suppl.), 366.

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

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in C: The art of scientific computing*. (2nd ed.) Cambridge: Cambridge University Press.

Strasburger, H. (2001a). Converting between measures of slope of the psychometric function. *Perception & Psychophysics*, **63**, 1348-1355.

Strasburger, H. (2001b). Invariance of the psychometric function for character recognition across the visual field. *Perception & Psychophysics*, **63**, 1356-1376.

Stromeyer, C. F., III, & Klein, S. A. (1974). Spatial frequency channels in human vision as asymmetric (edge) mechanisms. *Vision Research*, **14**, 1409-1420.

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

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

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

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

Wichmann, F. A., & Hill, N. J. (2001a). The psychometric function I: Fitting, sampling, and goodness-of-fit. *Perception & Psychophysics*, **63**, 1293-1313.

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

Yu, C., Klein, S. A., & Levi, D. M. (2001). *Psychophysical measurement and modeling of iso- and cross-surround modulation of the foveal TvC function*. Manuscript submitted for publication.

(*Continued on next page*)

APPENDIX
Matlab Programs; Available From klein@spectacle.berkeley.edu

Matlab Code 1
Weibull Function: Probability, d' and Log–Log d' Slope

Code for Generating Figure 1

```
'gamma = .5+.5*erf(-1/sqrt(2));                                              %gamma (lower asymptote) for z-score = -1
beta = 2; inc = .01;                                                         %beta is PF slope
x = inc/2:inc:3;                                                             %the stimulus range with linear abscissa
p = gamma+(1-gamma)*(1-exp(-x.^(beta)));                                     %Weibull function
subplot(3,2,1);plot(x,p);grid;text(.1,.92,'(a)'); ylabel('Weibull with beta = 2');
z = sqrt(2)*erfinv(2*p-1);                                                   %converting probability to z-score
subplot(3,2,3);plot(x,z);grid;text(.1,3.6,'(b)'); ylabel('z-score of Weibull (d´'-1)');
dprime = z+1;                                                                %dprime = z(hit) - z(false alarm)
difd = diff(log(dprime))/inc;                                                %derivative of log d'
xd = inc:inc:2.9999;                                                         %x values at midpoint of previous x values
subplot(3,2,5);plot(xd,difd.*xd);grid                                        %multiplication by xd is to make log abscissa
axis([0 3 .5 2]);text(.3,1.9,'(c)'); ylabel('log–log slope of d´''); xlabel('stimulus in threshold units')

y = -2:inc:1;                                                                %do similar plots for a log abscissa (y)
p = gamma+(1-gamma)*(1-exp(-exp(beta*y)));                                   %Weibull function as a function of y
subplot(3,2,2);plot(y,p,0,p(201),'*');grid; text(-1.8,.92,'(d)')
z = sqrt(2)*erfinv(2*p-1);
subplot(3,2,4);plot(y,z);grid;text(-1.8,3.6,'(e)')
dprime = z+1;difd = diff(log(dprime))/inc;
subplot(3,2,6);plot(y,[difd(1) difd]);grid; xlabel('stimulus in natural log units'); text(-1.6,1.9,'(f)')
```

Matlab Code 2
Goodness-of-Fit: Likelihood vs. Chi Square

Relevant to Wichmann and Hill (2001a) and Figure 4

```
clear, clf                                                                   %clear all
fac = [1 cumprod(1:20)];                                                     %create factorial function
p = [.51:.01:.999];                                                          %examine a wide range of probabilities
Nall = [1 2 4 8 16];                                                         %number of trials at each level
for iN = 1:5                                                                 %iterate over then number of trials
 N = Nall(iN); E = N*p;                                                      %E is the expected number as function of p
 lik = 0; X2 = 0;                                                            %initialize the likelihood and X^2
 for Obs = 0:N                                                               %sum over all possible correct responses
  Nm = N-Obs;                                                                %number of incorrect responses
  weight = fac(1+N)/fac(1+Obs)/fac(1+Nm)*p.^Obs.*(1-p).^Nm;                  %binomial weight
  lik = lik -2*weight.*(Obs*log(E/(Obs+eps))+ Nm*log((N-E)/(Nm+eps)));       %Equation 56
  X2 = X2+weight.*(Obs-E).^2./(E.*(1-p));                                    %calculate X^2, Equation 52
 end                                                                         %end summation loop
```

```
  LL2(iN,:) = lik;X2all(iN,:) = X2;                                                %store the likelihoods and chi squares
end                                                                                %end N loop
                                                                                   %the rest is for plotting
plot(p,LL2,'k',p,X2all(2,:),'--k');grid                                            %plot the log likelihoods and X^2
for in = 1:4;text(.935, LL2(in,end-6 ), num2str(Nall(in)));end
text(.913,1.2,'N = 16');text(.65,.95,'X2 for all N')
xlabel('P (predicted probability)')
ylabel('Contribution to log likelihood (D) from each level')
```

Matlab Code 3
Downward Slope Bias Due to Lapses

Relevant to Wichmann and Hill (2001a) and Table 2

```
function sse = probitML2(params, i,ilapse)                                         %function called by main program
stimulus = [0 1 6]; N = [50 50 50]; Obs = [25 42 50];                              %psychometric function information
lapseAll = [.01 .0001 .01 .0001];lapse = lapseAll(ilapse);                         %choose the lapse rate, lambda
if ilapse>2,Obs(3) = Obs(3)-1;end                                                  %produce a lapse for last 2 columns
zz = (stimulus-params(1))*params(2);                                               %z-score of psychometric function
ii = mod(i,3);                                                                     %index for selecting psychometric function
if ii = = 0,  prob = (1+erf(zz/sqrt(2)))/2;                                        %probit (cumulative normal)
elseif ii = = 1, prob = 1./(1+exp(-zz));                                           %logit
else   prob = 1-exp(-exp(zz));                                                     %Weibull
end
Exp = N.*(lapse+prob*(1-2*lapse));                                                 %expected number correct
if i<3, chisq = -2*(Obs.*log(Exp./Obs) + (N-Obs).*log((N-Exp+eps)./(N-Obs+eps)));  %log likelihood
else chisq = (Obs-Exp).^2 ./Exp./(N-Exp+eps);                                      %X^2, eps is smallest Matlab number
end
sse = sum(chisq);                                                                  %sum over the three levels

%%**MAIN PROGRAM**
clear;clf
type = ['probit maxlik ';'logit maxlik ';'Weibull maxlik ';                        %headings for Table 2
  'probit chisq ';'logit chisq ';'Weibull chisq ';];
disp('  gamma =  .01  .0001  .01  .0001')                                          %type top of Table 2
disp('  lapse =  no  no   yes   yes')
for i = 0:5                                                                        %iterate over function type
  for ilapse = 1:4                                                                 %iterate over lapse categories
params = fmins('probitML2',[.1 .3],[],[],i,ilapse);                                %fmins finds minimum of chi-square
slope(ilapse) = params(2);                                                         %slope is 2nd parameter
  end
  disp( [type(i+1,:) num2str(slope)] )                                             %print result
end
```

APPENDIX (Continued)

Matlab Code 4
Brownian Staircase Enumerations for Slope Bias

Monotony

```
flag = 1;ii = 0;                                            %initialize, flag = 1 means nonmonotonicity is detected
while flag = = 1,                                           %keep going if there is nonmonotonic data
  flag = 0;                                                 %reset monotonic flag
  ii = ii+1;                                                %increase the nonmonotonic spread
  for i2 = ii+1:ntot;                                       %loop over all PF levels looking for nonmonotonicity
    prob = cor./(tot+eps);                                  %calculate probabilities
    if (prob(i2-ii)>prob(i2)) & (tot(i2)>0),                %check for nonmonotonicity
      cor(i2-ii:i2) = mean(cor(i2-ii:i2))*ones(1,ii+1);     %replace nonmonotonic correct by average
      tot(i2-ii:i2) = mean(tot(i2-ii:i2))*ones(1,ii+1);     %replace nonmonotonic total by average
      flag = 1;                                             %set flag that nonmonotonicity is found
end; end; end
```

Note that in line 6 the denominator is tot+eps, where eps is the smallest floating point number that Matlab can use. Because of this choice, if there are zero trials at a level, the associated probability will be zero.

Extrapolate

```
ii = 1; while tot(ii) = = 0;ii = ii+1;end                   %count the low levels with no trials
if ii>1,                                                    %skip lowest level
  tot(1:ii-1) = ones(1,ii-1)*.00001;                        %fill in with very low number of trials
  cor(1:ii-1) = prob(ii)*tot(1:ii-1);                       %fill in with probability of lowest tested level
end
ii = nbits2; while tot(ii) = = 0;ii = ii-1;end              %do the same extrapolation at high end
if ii<nbits2,
  tot(ii+1:nbits2) = ones(1,nbits2-ii)*.00001;
  cor(ii+1:nbits2) = prob(ii)*tot(ii+1:nbits2);
end
```

Averaging

```
prob1 = cor./(tot+eps);                                     %calculate PF
probAll(iopt,:) = probAll(iopt,:)+prob1;                    %sum the PFs to be averaged
probTotAll(iopt,:) = probTotAll(iopt,:)+(tot>0);            %count the number of staircases with a given level
corAll(iopt,:) = corAll(iopt,:)+cor;                        %sum the number correct over all staircases
totAll(iopt,:) = totAll(iopt,:)+tot;                        %sum the total number of trials over all staircases
```

Main Program

```
nbits = 10;n2 = 2^nbits-1;nbits2 = 2*nbits-1;               %nbits is number of staircase steps
zero3 = zeros(3,nbits2);                                    %the three rows are for the three iopt values
for ishift = 0:1                                            %ishift = 0 means no shift (1 means shift)
totAll = zero3; corAll = zero3; probAll = zero3;probTotAll = zero3;   %initialize arrays
```

```
  for i = 0:n2                                          %loop over all possible staircases
  a = bitget(i,[1:nbits]);                              %a(k) = 1 is hit, a(k) = 0 is miss
  step = 1-2*a(1:end-1);                                %1 step down for hit, 1 step up for hit
  aCum = [0 cumsum(step)];                              %accumulate steps to get stimulus level
  if ishift = = 0, ave = 0;                             %for the no-shift option
  else ave = round(mean(aCum));end                      %for shift option, ave is the amount of shift
  aCum = aCum-ave+nbits;                                %do the shift
  cor = zeros(1,nbits2); tot = cor;                     %initialize for each staircase
  for i2 = 1:nbits;
  cor(aCum(i2)) = cor(aCum(i2))+a(i2);                  %count number correct at each level
  tot(aCum(i2)) = tot(aCum(i2))+1;                      %count number trials at each level
    end
  iopt = 1;stairave;                                    %two types of averaging (iopt is index in script above)
  iopt = 2;monotonize2;stairave;                        %monotonize PF and then do averaging
  iopt = 3;extrapolate;stairave;                        %extrapolate and then do averaging
end
prob = corAll./(totAll+eps);                            %get average from pooled data
probave = probAll./(probTotAll+eps);                    %get average from individual PFs
x = [1:nbits2];                                         %x axis for plotting
subplot(3,2,1+ishift);                                  %plot the top pair of panels
plot(x,prob(1,:),x,totAll(1,:)/max(totAll(1,:)),'-.',x,probave(1,:),'--');grid
if ishift = = 0,title('no shift');ylabel('no data manipulation');text(.5,.95,'(a)')
else title('shift');text(.5,.95,'(d)');end
subplot(3,2,3+ishift);
plot(x,prob(2,:),x,probave(2,:),'--');grid;             %plot middle pair of panels
if ishift = = 0, ylabel('monotonize psychometric fn');text(.5,.63,'(b)');
     else text(.5,.95,'(e)');end
subplot(3,2,5+ishift);                                  %plot bottom pair of panels
plot(x,prob(3,:),x,probave(3,:),'--',[nbits nbits],[.45 .55]);grid
tt = 'cf';text(.5,.95,['(' tt(ishift+1) ')'])
if ishift = = 0, ylabel('extrapolate psychometric fn');end
xlabel('stimulus levels')
end                                                     %end loop of whether to shift or not
```

Note—The main program has one tricky step that requires knowledge of Matlab: a = *bitget*(i,[1:nbits]);. where *i* is an integer from 0 to $2^{\text{nbits}}-1$, and nbits = 10 in the present program. The *bitget* command converts the integer *i* to a vector of bits. For example, for *i* = 11 the output of the *bitget* command is 1 1 0 1 0 0 0 0 0 0. The number *i* = 11 (binary is 1011) corresponds to the 10 trial staircase C C I C I I I I I I, where C means correct and I means incorrect.

(Manuscript received July 10, 2001;
accepted for publication August 8, 2001.)