Perception & Psychophysics
2001, 63 (8), 1399-1420

On the analysis of psychometric functions: The Spearman–Kärber method

JEFF MILLER
University of Otago, Dunedin, New Zealand
and
ROLF ULRICH
University of Tübingen, Tübingen, Germany

With computer simulations, we examined the performance of the Spearman–Kärber method for analyzing psychometric functions and compared this method with the standard technique of probit analysis. The Spearman–Kärber method was found to be superior in most cases. It generally yielded less biased and less variable estimates of the location and dispersion of a psychometric function, and it provided more power to detect differences in these parameters across experimental conditions. Moreover, the Spearman–Kärber method provided information about the skewness and higher moments of psychometric functions that is beyond the scope of probit analysis. These advantages of the Spearman–Kärber method suggest that it should often be used in preference to probit analysis for the analysis of observed psychometric functions.

One of the most important tools in psychophysics is the psychometric function. This function can be used to estimate both absolute thresholds and difference limens (*dl*s) (e.g., Gescheider, 1997; Woodworth & Schlosberg, 1954), and it is also sometimes used to test specific psychophysical models (e.g., Falmagne, 1985; Sternberg & Knoll, 1973). An observed psychometric function plots the proportion of times a certain response is given as a function of some property of a physical stimulus. For example, Figure 1 shows a psychometric function observed in a duration discrimination task (Getty, 1975). In this task, an observer was presented sequentially with a standard tone of duration d_s and a comparison tone of duration d_c, and the observer reported whether the comparison tone was longer than the standard. For any standard duration d_s, the psychometric function shows the probability of the "comparison longer" response as a function of the duration of the comparison tone. As one would expect, the function increases with the duration of the comparison tone.

Two measures computed from an observed psychometric function are usually of particular interest. First, the experimental focus is often on the steepness of the psychometric function, most commonly measured using the *dl*. The *dl* is usually defined as half of the interquartile range of this function, which means that more precise judgments yield smaller values of *dl*. Second, the location of the psychometric function along the physical stimulus dimension is often of interest, to assess whether a certain experimental manipulation affects perception (e.g., Flanagan & Wing, 1997; Mattes & Ulrich, 1998). In discrimination tasks, the location of the psychometric function is usually measured with the point of subjective equality (*pse*), which corresponds to the stimulus point at which a certain response is given in 50% of all trials. In detection tasks, the 50% point is sometimes used to define the absolute stimulus threshold.[1]

In both discrimination and detection tasks, the psychometric function can be characterized more generally as a function relating the probability of a certain response, r, to the value of a stimulus, s, along a certain physical dimension:

$$F(s) = Pr\{R = r \mid S = s\}. \quad (1)$$

It is useful to distinguish between true and observed psychometric functions. A true underlying function must satisfy three conditions: (1) $F(s) \to 0$ as $s \to -\infty$, (2) $F(s) \to 1$ as $s \to \infty$, and (3) $F(s)$ is monotonically increasing with s (Falmagne, 1985; Luce, 1963; Urban, 1907). The true function is, of course, unknown in any realistic experimental situation, and the goal of the experiment is to investigate it. In contrast, an observed psychometric function consists of a relatively small set of estimated response probabilities at distinct stimulus values. Crucially, as will be discussed in detail later, the estimated probabilities need not be monotonic, because they are obtained from a set of experimental trials and are thus subject to binomial random error.

Under the three conditions assumed to characterize true functions, any psychometric function $F(s)$ can be regarded

This work was supported by cooperative research funds from the Deutsche Raum- und Luftfahrtgesellschaft e.V. The authors thank Stanley Klein for helpful comments on the manuscript. Correspondence concerning this article should be addressed to J. Miller, Department of Psychology, University of Otago, Dunedin, New Zealand, or Rolf Ulrich, Abteilung für Allgemeine Psychologie und Methodenlehre, Psychologisches Institut, Universität Tübingen, Friedrichstr. 21, 72072 Tübingen, Germany (e-mail: miller@otago.ac.nz or ulrich@uni-tuebingen.de).

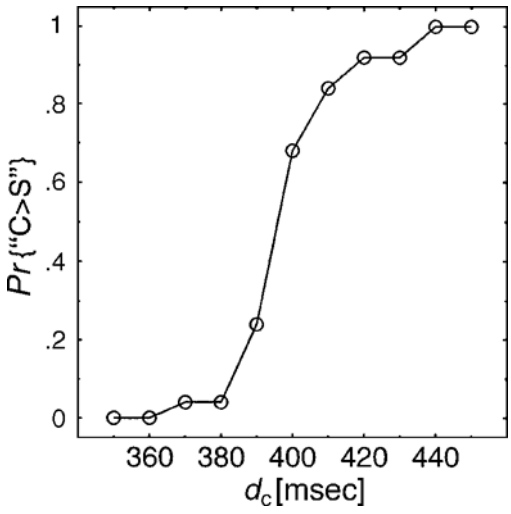

Figure 1. Data from a duration discrimination experiment of Getty (1975). The function shows the observed proportion of trials in which a comparison stimulus (C) was judged to be longer than a standard stimulus (S), *Pr*{"C > S"}, as a function of the duration of the comparison stimulus, d_c. These data depict the psychometric function obtained for the observer D.G., using a standard duration of 400 msec.

as the cumulative density function (CDF) from some probability distribution (Trevan, 1927), and it is very convenient to do so for at least two reasons. First, most models used to predict psychometric functions assume that each stimulus elicits a sensory magnitude that varies from trial to trial according to some probability distribution and that this magnitude is compared against an internal criterion in order to select the response (see Falmagne, 1985, for a rigorous treatment of psychophysical models of both detection and discrimination). According to such models, the true value of the psychometric function at each stimulus value is the probability that the evoked sensory magnitude exceeds the decision criterion. For example, the well-known threshold theory (i.e., the phi–gamma hypothesis) assumes that a stimulus of a particular intensity is detected only on those trials in which the stimulus exceeds a momentary threshold (see Gescheider, 1997; Woodworth & Schlosberg, 1954, p. 221). Because factors affecting this threshold fluctuate randomly from moment to moment, the same stimulus will be detected on some trials, but not on others. For example, if the momentary threshold is normally distributed with mean μ and standard deviation σ, the psychometric function corresponds to the CDF of a normal distribution with these parameters.

Second, when psychometric functions are regarded as CDFs, they can be conveniently summarized in terms of the parameters of the underlying probability distribution. Most commonly, the location of the psychometric function is summarized in terms of the distribution's mean, median, or both, and its dispersion is summarized by the distribution's standard deviation or interquartile range. For the psychometric function depicted in Figure 1, for example, probit analysis (discussed further below) provides estimates of the mean and standard deviation of the underlying distribution of $\hat{\mu}_{PR}$ = 384 msec and $\hat{\sigma}_{PR}$ = 10 msec, respectively, where the subscript PR indicates that estimates were computed using the probit method, as opposed to the alternative Spearman–Kärber (SK) method, with which it is compared in this article. In addition, our notation will henceforth distinguish between true parameters of psychometric functions written without hats, such as μ and σ, and estimates of those parameters written with hats, such as $\hat{\mu}_{PR}$ and $\hat{\sigma}_{PR}$.

Given that a true psychometric function $F(s)$ cannot be observed directly but can only be inferred from the results of experimental trials, statistical methods are needed to estimate the true psychometric function from the observed one. The classical (e.g., Fechner, 1860) and most common procedure for doing this is probit analysis (e.g., Finney, 1952, 1978). In this type of analysis, the psychometric function is assumed to have the shape of a cumulative normal distribution, often referred to as a normal ogive. Given this assumption, the data are used to estimate the parameters of the normal distribution (i.e., μ and σ) via one of several alternative techniques (e.g., Finney, 1978; Foster & Bischof, 1987).

In a variety of situations, however, probit analysis does not seem completely appropriate. One problem with this type of analysis is that there are often good reasons to suppose that the underlying distribution is not normal, and in these cases the use of the inappropriate normal measurement model may distort estimates of the location, dispersion, and other parameters of the psychometric function (e.g., Falmagne, 1985; Mortensen, 1988; Quick, 1974; Sternberg & Knoll, 1973). For example, the fit of an observed psychometric function to the normal can be evaluated with a χ^2 test (Guilford, 1936), and the results often allow rejection of the normal model (e.g., G. A. Miller & Garner, 1944; Stevens, Morgan, & Volkmann, 1941). As another example, Weber's law provides a very general argument that true psychometric functions should be positively skewed, rather than symmetric, because a given change in stimulus value should have a greater effect at the bottom of the stimulus range than at the top (Guilford, 1954; but see also Falmagne, 1985). Therefore, the generality of Weber's law suggests that the normal may often be inappropriate as a model of the psychometric function. Consistent with this argument, we reanalyzed the data of Getty (1975) and found evidence that the observed psychometric functions are positively skewed, rather than symmetric [$t(29) = 3.97$, $p < .001$].[2] Moreover, in specific situations, there are sometimes good theoretical arguments suggesting that the psychometric function has a nonnormal shape. For example, the physics of quantal fluctuations suggest that psychometric functions for detection of weak visual stimuli should be determined by a Poisson process under some circumstances (e.g., Gescheider, 1997, chap. 4), and analyses of the energy in acoustic signals and signal-plus-noise segments suggest that psychometric functions for the detection of sinusoidal signals in noise should have the shape of a noncentral F distribution (Green & McGill, 1970). Similarly, models involving temporal probability summa-

tion among independent subsystems predict Weibull functions (e.g., Green & Luce, 1975; Nachmias, 1981; Quick, 1974), and peak-detection models predict asymmetric psychometric functions (e.g., Mortensen, 1988), as does the pseudologistic model of timing suggested by Killeen, Fetterman, and Bizo (1997). In some cases, it may be possible to transform the stimulus values to normalize the psychometric function, as required by probit analysis (e.g., via the log transform; Finney, 1952), but there is no guarantee that this approach will always be successful.

A second problem with probit analysis is that, in many cases, the psychophysical theories being tested make no specific assumptions about the underlying distributional shape and thus generate distribution-free predictions. In these cases, it is desirable—sometimes, even crucial—to examine the predictions of the theories without imposing such assumptions in order to estimate parameters. For example, some psychophysical theories assume that responses are determined by the sums of internal random variables, and it is possible to derive predictions about the moments (e.g., mean, *SD*, skewness) of psychometric functions, but not about percentile-based measures (e.g., median or *pse*, *dl*) or distributional shapes (e.g., Sternberg & Knoll, 1973; Sternberg, Knoll, & Zukofsky, 1982; Ulrich, 1987). The assumption of normality is gratuitous with respect to tests of predictions concerning means and standard deviations, and worse yet it precludes tests of predictions concerning skewness and higher moments. As another example, some theories predict that two or more observed psychometric functions should be parallel (e.g., Allan, 1975; Falmagne, 1985; Green & Luce, 1975; Mortensen & Suhl, 1991; Ulrich, 1987). This prediction can be supported if the higher moments of these functions are identical (see Ulrich, 1987), but a complete test of this prediction requires the estimation of higher moments without assuming a specific functional form of the psychometric functions. Indeed, if the normal shape is assumed, the psychometric functions can only be compared with respect to their dispersions, because fitting a normal ogive to each condition ensures equality of the estimated third and higher moments. In sum, then, it would clearly be desirable in a number of situations to have an alternative to probit analysis that does not require any specific assumptions concerning the distributional shape of the true psychometric function.

Although it is little used, the SK method does provide an alternative distribution-free method for the estimation of psychometric functions (e.g., Epstein & Churchman, 1944; Kärber, 1931; Spearman, 1908). As will be described further below, this method can be used to estimate any desired parameters of a psychometric function, including not only its location and dispersion, as provided by probit analysis, but also its skewness, kurtosis, and so on. Because it is distribution free (i.e., it makes weaker assumptions about the true psychometric function), the SK method may be preferred to probit analysis in situations in which one is reluctant to assume a specific functional form for the true psychometric function.

The SK method has at least two potential practical advantages, as compared with probit analysis. First, it may provide more accurate estimates of location and dispersion parameters, especially when the assumption of normality is violated. Second, the additional parameters it can estimate (e.g., skewness) may be useful not only for describing psychometric functions, but also for testing specific models that make predictions about these parameters, without making any assumptions about distributional shapes.

In this article, we report computer simulations comparing probit analysis of psychometric functions with analysis using the SK method. We first describe the latter method in more detail and then present simulations evaluating and comparing the two methods. Our simulations address a complex and interrelated set of questions that are of practical interest to experimenters who wish to estimate the parameters of psychometric functions. In separate sections, we consider estimators of four types of parameters that might be of interest: the position of the psychometric function along the stimulus axis (i.e., location), its steepness (i.e., dispersion), its deviation from symmetry (i.e., skewness), and its relative peakedness (i.e., kurtosis). For each estimator, we consider several questions, including the following. (1) How biased is it? (2) What is its standard error? (3) How likely is it that confidence intervals computed around it actually contain the true parameter value? and (4) How much power does it have to detect differences between experimental conditions? Thus, these simulations extend previous biometrical work examining the properties of means estimated with the SK method (see Cornell, 1983), as well as providing a direct comparison with estimates obtained using the probit method.

The Spearman–Kärber Method

The SK method treats the observed psychometric function as a cumulative distribution function for grouped data from which the corresponding histogram is reconstructed. Specifically, suppose that an experimenter uses k stimulus values, $s_1 < s_2 \ldots < s_k$, to determine the observed response probability, $\hat{p}_i$, ($i = 1, \ldots, k$), associated with each stimulus value. The estimated probability associated with the stimulus range (or bin) from s_{i-1} to s_i is then equal to $\hat{p}_i - \hat{p}_{i-1}$, and assuming a uniform distribution within the bin, the probability density within the bin is thus estimated by $(\hat{p}_i - \hat{p}_{i-1})/(s_i - s_{i-1})$. Thus, this histogram can be used to approximate the continuous distribution that underlies the data.

With this conception of the underlying probability distribution, Sternberg et al. (1982, pp. 234–236; cf. Church & Cobb, 1973) provided a modified SK method to estimate the rth raw moment μ'_r of the psychometric function. Specifically, the estimated rth raw moment is given by

$$\hat{\mu}'_r = \frac{1}{r+1} \sum_{i=1}^{k+1} \frac{\left(\hat{p}_i - \hat{p}_{i-1}\right)\left(s_i^{r+1} - s_{i-1}^{r+1}\right)}{s_i - s_{i-1}}. \qquad (2)$$

The stimulus levels s_0 ($s_0 < s_1$) and s_{k+1} ($s_{k+1} > s_k$) are chosen such that one can assume true values of $p_0 = 0$ and $p_{k+1} = 1$. Note that the specification of these two extreme values is necessary only if there is a truncation

error—that is, if the stimulus series $s = (s_1, \ldots, s_k)$ is not broad enough to cover the whole transition zone of the true psychometric function (i.e., $\hat{p}_1 > 0$ or $\hat{p}_k < 1$; see Woodworth & Schlosberg, 1954, p. 209). Thus, although many stimulus levels are not necessarily required, the interval between the levels should be large enough to cover the whole transition zone.

As a numerical illustration of Equation 2, consider the response probabilities $\hat{p} = (.0, .03, .15, .55, .87, 1.0)$ obtained at the six stimulus levels $s = (3, 6, 9, 12, 15, 18)$. Because $\hat{p}_1 = 0$ and $\hat{p}_k = 1$, there is no truncation error, and the values of s_0 and s_{k+1} do not affect the estimates. Thus, we can arbitrarily set s_0 to 1 and s_{k+1} to 20. For this example, the first four estimated raw moments are $\hat{\mu}'_1 = 11.70$ (i.e., the arithmetic mean), $\hat{\mu}'_2 = 145.92$, $\hat{\mu}'_3 = 1{,}914.03$, and $\hat{\mu}'_4 = 26{,}172.72$. Thus, for this example the second, third, and fourth moments about the mean (i.e., variance, skewness, and kurtosis; see Stuart & Ord, 1987, p. 73)

$$\hat{\mu}_2 = \hat{\mu}'_2 - \left(\hat{\mu}'_1\right)^2 = 9.03, \tag{3}$$

$$\hat{\mu}_3 = \hat{\mu}'_3 - 3\hat{\mu}'_2\hat{\mu}'_1 + 2\left(\hat{\mu}'_1\right)^3 = -4.54, \tag{4}$$

and

$$\hat{\mu}_4 = \hat{\mu}'_4 - 4\hat{\mu}'_3\hat{\mu}'_1 + 6\hat{\mu}'_2\left(\hat{\mu}'_1\right)^2 - 3\left(\hat{\mu}'_1\right)^4 = 229.43. \tag{5}$$

Note that the estimated third moment is negative, suggesting that the true function is negatively skewed. The estimate of the fourth moment can be used to assess the kurtosis of the underlying distribution. The sample kurtosis is sometimes expressed in the following standardized form (e.g., Stuart & Ord, 1987, p. 107):

$$\hat{\gamma}_2 = \frac{\hat{\mu}_4}{\left(\hat{\mu}_2\right)^2}. \tag{6}$$

The normal distribution yields $\gamma_2 = 3$. In contrast, distributions with $\gamma_2 > 3$ tend to have thick tails and be more peaked than the normal, whereas those with $\gamma_2 < 3$ tend to have thin tails and a relatively low peak (for further information, see DeCarlo, 1997). For our example, then, one computes

$$\hat{\gamma}_2 = \frac{229.43}{9.03^4} \approx 0.03, \tag{7}$$

and this value indicates that the psychometric function of our numerical example has rather thin tails and a low peak, relative to the normal.

The median, *dl*, and other indices based on the percentiles of the psychometric function can also be estimated using the SK method. Specifically, the estimated value at any percentile can be computed via linear interpolation from the obtained response probabilities. In the above numerical example, linear interpolation yields a median of 11.6 and a *dl* of 2.1.

Nonmonotonic Functions

A complication for the SK method is that observed psychometric functions can sometimes be nonmonotonic even though the true underlying psychometric function is nondecreasing. This is especially likely to happen in experiments in which there are many stimulus values and only a few trials per stimulus value, because in such experiments the binomial variability in estimated response probabilities is relatively large. When an observed psychometric function is nonmonotonic, it should be monotonized before the SK method can be applied, because certain parameters (e.g., variance, percentiles) of the true psychometric function are clearly better estimated from nondecreasing response probabilities than from nonmonotonic ones (Sternberg et al., 1982).

Ayer, Brunk, Ewing, Reid, and Silverman (1955) described an algorithm that can be used to monotonize an observed psychometric function when needed. Their algorithm provides a maximum likelihood estimate for the true response probability at each stimulus value. In brief, these maximum likelihood estimates $\bar{p} = (\bar{p}_1 \leq \ldots \leq \bar{p}_k)$ are computed from the obtained response frequencies $X = (X_1, \ldots, X_k)$, where X_i denotes the observed frequency of response r (e.g., "yes" responses) at stimulus level s_i when there are a total of n_i trials at this level (see Ayer et al., 1955, p. 641). This computation is actually intuitively appealing and will be subdivided into two steps.

Step 1. If $\hat{p}_1 \leq \hat{p}_2 \leq \ldots \hat{p}_k$, then $\bar{p}_i = \hat{p}_i$ $(i = 1, \ldots, k)$, and the estimation process is completed. Otherwise, proceed to Step 2.

Step 2. If two consecutive values $\hat{p}_i$ and $\hat{p}_{i+1}$ are nonmonotone—that is, $\hat{p}_i > \hat{p}_{i+1}$—then replace their ratios $\hat{p}_i = X_i/n_i$ and $\hat{p}_{i+1} = X_{i+1}/n_{i+1}$ by the single ratio $\hat{p}_i = \hat{p}_{i+1} = (X_i + X_{i+1})/(n_i + n_{i+1})$. Likewise, if three consecutive values $\hat{p}_i > \hat{p}_{i+1} > \hat{p}_{i+2}$ are nonmonotone, then replace the corresponding ratios by the single ratio $\hat{p}_i = \hat{p}_{i+1} = \hat{p}_{i+2} = (X_i + X_{i+1} + X_{i+2})/(n_i + n_{i+1} + n_{i+2})$. Should more than three consecutive values be nonmonotone, group their ratios analogously. Repeat Step 1.

Table 1 provides a numerical example illustrating this method. The top section of the table shows the raw data obtained from a hypothetical experiment. The first row shows the observed response probabilities that were computed from the observed frequencies (second row) and the number of trials at each stimulus level (third row). The second and third sections show the analogous quantities after the two needed steps in which nonmonotonic observed response probabilities were combined. The last row shows the final maximum likelihood estimates of the response probabilities after the monotonizing procedure has been completed.[3]

METHOD

We evaluated the SK method and compared it with probit analysis in simulations of experiments using the method of constant stimuli. Except as will be noted later, the different sets of simulations were defined by the factorial combination of (1) the number of stimulus levels selected by the experimenter (5, 11, or 21), (2) the number of trials used to estimate the psychometric function (approximately 30, 40, 60, 100, 300, or 1,000, as will be ex-

Table 1
Illustration of Ayer, Brunk, Ewing, Reid, and Silverman's (1955) Method for Maximum Likelihood Estimation of True Response Probabilities When Observed Response Probabilities are Nonmonotonic

Measure	Stimulus Level s_1	s_2	s_3	s_4	s_5
		Raw Data			
$\hat{p}_i$	.30	.28	.20	.90	.53
X_i	15	28	8	54	16
n_i	50	100	40	60	30
		Combining Levels 1–3			
$\hat{p}_i$	.27	.27	.27	.90	.53
X_i		51		54	16
n_i		190		60	30
		Combining Levels 4 and 5			
$\hat{p}_i$	.27	.27	.27	.78	.78
X_i		51		70	
n_i		190		90	
		Final Estimates			
$\bar{p}_i$	.27	.27	.27	.78	.78

Note—In this example, there are five stimulus levels, s_1 to s_5. For each stimulus level, the observed numbers of trials and r responses are n_i and X_i, respectively, and the estimated probability of response r is $\hat{p}_i$. Using Ayer et al.'s (1955) method, stimulus levels are combined as illustrated to eliminate nonmonotonicity in the values of $\hat{p}_i$. The final maximum likelihood estimates, $\bar{p}_i$, are those obtained after all nonmonotonicities have been removed.

plained later), and (3) the true underlying psychometric function, which is one of the 20 possibilities listed in Table 3 and will be described in detail below. Thirty thousand experiments were simulated within each set, and the results were summarized across these experiments.

Simulation Protocol

It is convenient to first describe in detail the procedure for simulating one example experiment and later to describe other simulations by indicating their differences from it. The example to be described came from the set using 11 stimulus levels, approximately 300 trials, and an underlying psychometric function having the shape of a cumulative normal distribution with a true mean of $\mu = 0.5$ and standard deviation of $\sigma = 0.25$. The stimulus levels were equally spaced, and the smallest and largest stimulus levels were chosen to be the 1st and 99th percentiles of the true psychometric function, respectively. These constraints uniquely determine the stimulus values to be those shown in the top row of Table 2. The true probability of the response r at each stimulus level could then be computed as the CDF of the true underlying distribution at that stimulus level, and the probabilities for this example are also shown in the second row of Table 2.

To simulate an experiment, the 300 experimental trials were divided as equally as possible across the 11 stimulus levels, subject to the constraint of equal numbers of trials per stimulus, so there were 27 trials per stimulus in these simulations.[4] Each of these trials was simulated by generating a uniform random number in the interval 0–1. If this random number was less than the true response probability, we generated the response r for that trial; otherwise, we generated the alternative response $\bar{r}$. We counted the numbers of r and $\bar{r}$ responses at each stimulus level just as an experimenter would for real observers, and example data for one simulation are shown in Table 2.

This simulation protocol varied in obvious ways as the simulation conditions were changed across simulation sets. For example, the number of simulated trials for each stimulus level depended on the number of experimental trials in that simulation set (i.e., 30, 40, 60, 100, 300, or 1,000). Similarly, the number of stimulus levels also changed across simulation sets (5, 11, or 21), thereby allowing comparison of simulations in which the psychometric function was sampled more or less sparsely. In addition, we varied the true underlying psychometric function across simulation sets, using 20 different true functions, as will be described in the next section. The true function determined the values of the smallest and largest stimulus values because these were always placed at the 1st and 99th percentiles of the true distribution, respectively.[5] These values, in conjunction with the number of stimulus levels in the simulation and the requirement of equal spacing, determined the intermediate stimulus values as well. Finally, the true function determined the true probabilities of the two responses at each stimulus value.

True Psychometric Functions

Table 3 lists the true underlying psychometric functions used in the different sets of simulations and the pa-

Table 2
Stimulus Values, True Probabilities p_i of Response r, and Example Simulated Results for Simulations With Normal Psychometric Function, 11 Stimulus Levels, and 300 Experimental Trials

	Stimulus Level 1	2	3	4	5	6	7	8	9	10	11
All simulations											
Stimulus value s_i	−0.08	0.04	0.15	0.27	0.38	0.50	0.62	0.73	0.85	0.97	1.08
True probability p_i	.01	.03	.08	.18	.32	.50	.68	.82	.92	.97	.99
One simulation											
N of trials with response r	0	2	2	61	10	12	16	25	26	26	27
N of trials with response $\bar{r}$	27	25	25	21	17	15	11	2	1	1	0
Estimated probability $\hat{p}_i$	.0	.07	.07	.22	.37	.44	.59	.93	.96	.96	1.0

rameter values of these functions (i.e., μ, σ, etc.). The individual functions are briefly described next; Figure 2 shows plots of them, and the Appendix gives the exact equations for them. Note that it was not possible to match all the functions exactly with respect to their means and standard deviations (e.g., the quantal and Naka–Rushton distributions cannot be matched on these parameters). Where appropriate, then, we corrected simulation results for inherent differences in distributional properties before combining across distributions.

Normal distribution. Of course, the normal distribution was used in one set of simulations as a convenient baseline distribution (see Simpson, 1995). In addition, because it is the underlying distribution assumed by probit analysis, a comparison of the probit and SK methods with this distribution can indicate whether there are costs of using the distribution-free analysis when the distributional assumption is in fact satisfied.

Quantal distribution. The quantal distribution is also theoretically motivated. As has been discussed by Gescheider (1997, pp. 81–86), this distribution represents the predicted psychometric function when sensitivity is determined by the number of quanta emitted by a Poisson process with a mean equal to the stimulus value. A parameter of this distribution is the criterion number of quanta needed for response r to occur, and we chose a criterion of four to obtain a relatively skewed quantal distribution.

Naka–Rushton and Weibull distributions. The Naka–Rushton and Weibull distributions are also reasonable distributions on theoretical grounds and have been used for previous simulations of psychometric functions (e.g., Simpson, 1995). Both are skewed and rather nonnormal in shape.[6]

Model 4 distribution. The model 4 distribution is a rather unusual distribution that arises from a particular model of temporal order judgment considered by Sternberg and Knoll (1973). This distribution was selected because of its large deviation from the normal distribution. It consists of two nonoverlapping uniform (i.e., rectangular distribution) segments, one from $-t$ to 0 and the other from t to $2t$. We used the version of the distribution with $t = 1$.

Triangular distributions. We also included five triangular distributions defined over stimulus values in the range of 0–1. The triangular distribution arises as a model of the psychometric function in human duration discrimination (Kristofferson, 1984). In addition, it is convenient to use the triangular distribution because its skew can be adjusted, thereby allowing us to compare the effectiveness of the probit and SK analyses with different amounts of skew in the underlying distribution. We used triangular distributions with skewness values of −0.8, −0.4, 0, 0.4, and 0.8.[7]

Mixture distributions. We also included five mixture distributions defined over stimulus values in the range of 0–1. Each was the equally weighted mixture of a uniform distribution over this whole range and a second uniform at one end of the range (i.e., 0–a or a–1, where a is a constant chosen individually for each distribution). Although these mixture distributions are not easily motivated theoretically, they also allow convenient adjustment of skew by variation of a, thereby providing an addi-

Table 3
The Cumulative Probability Distributions Used as Models of the True Underlying Psychometric Function and the Parameters of Each Distribution

	Parameter							
Distribution	μ	*pse*	σ	*dl*	$\gamma_1(M)$	$\gamma_1(P)$	$\gamma_2(M)$	$\gamma_2(P)$
Normal	0.50	0.50	0.25	0.17	0.00	0.00	3.00	0.26
Quantal	4.00	3.67	2.00	1.29	1.00	0.12	4.49	0.26
Naka–Rushton	1.47	1.00	1.68	0.57	1.61	0.27	28.79	0.22
Weibull	0.89	0.83	0.46	0.32	0.86	0.08	3.24	0.27
Model 4	0.50	0.50	1.04	1.00	0.00	0.00	1.29	0.38
Triangular(−0.8)	0.61	0.65	0.22	0.17	−0.80	−0.13	2.40	0.29
Triangular(−0.4)	0.51	0.52	0.20	0.15	−0.40	−0.03	2.40	0.27
Triangular(0.0)	0.50	0.50	0.20	0.15	0.00	0.00	2.40	0.26
Triangular(0.4)	0.49	0.48	0.20	0.15	0.40	0.03	2.40	0.27
Triangular(0.8)	0.39	0.35	0.22	0.17	0.80	0.13	2.40	0.29
Mixture(−0.8)	0.61	0.63	0.26	0.18	−0.80	−0.00	2.42	0.25
Mixture(−0.4)	0.54	0.54	0.27	0.23	−0.40	0.00	1.89	0.31
Mixture(0)	0.50	0.50	0.29	0.25	0.00	0.00	1.80	0.31
Mixture(0.4)	0.46	0.46	0.27	0.23	0.40	0.00	1.89	0.31
Mixture(0.8)	0.39	0.37	0.26	0.18	0.80	−0.00	2.42	0.25
$t(5)$	0.00	0.00	1.29	0.73	0.00	−0.00	7.81	0.25
$t(6)$	0.00	0.00	1.22	0.72	0.00	−0.00	5.78	0.25
$t(7)$	0.00	−0.00	1.18	0.71	0.00	0.00	4.92	0.25
$t(10)$	0.00	0.00	1.12	0.70	0.00	−0.00	3.99	0.25
$t(16)$	0.00	0.00	1.07	0.69	0.00	−0.00	3.50	0.26

Note—μ = mean, *pse* = point of subjective equality (median), σ = standard deviation, *dl* = difference limen, $\gamma_1(M)$ = moment-based measure of skewness $(E[(x-\mu)^3])/(\sigma^3)$, $\gamma_1(P)$ = percentile-based measure of skewness $(s_{75} - 2 \cdot s_{50} + s_{25})/(s_{75} - s_{25})$, $\gamma_2(M)$ = moment-based measure of kurtosis $(E[(x-\mu)^4])/(\sigma^4)$, and $\gamma_2(P)$ = percentile-based measure of kurtosis $(dl)/(s_{90} - s_{10})$, where s_p is the stimulus value at the pth percentile of the distribution. The skewness and kurtosis measures are discussed in detail in later sections. The parameters of the triangular and mixture distributions are their skewness values, and the parameters of the t distributions are their degrees of freedom values.

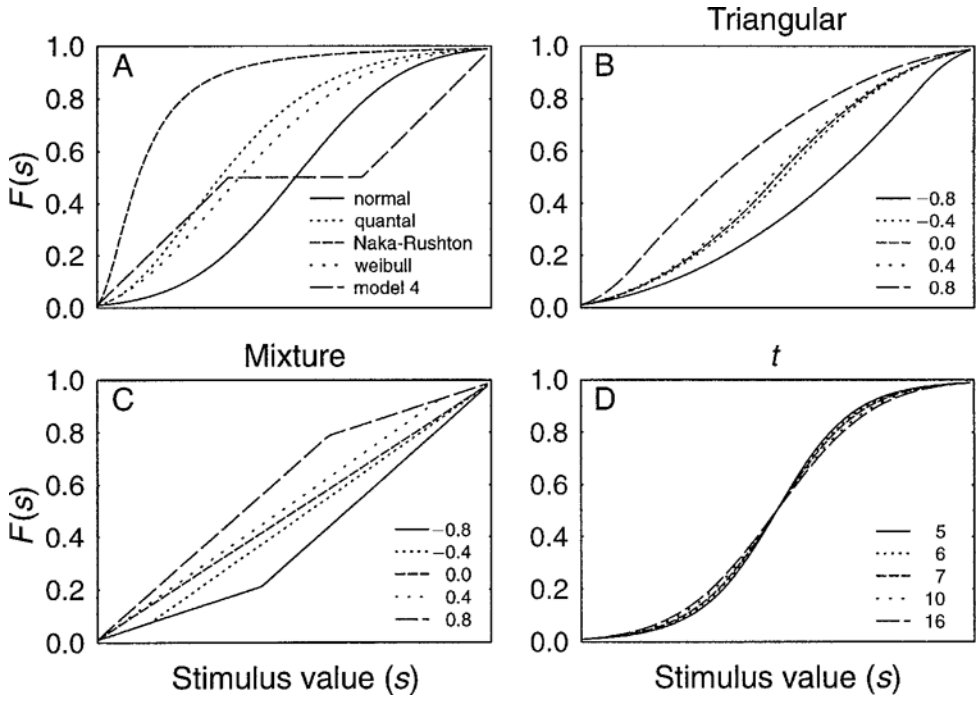

Figure 2. Plots of the true psychometric functions listed in Table 3 and used in the simulations. Panel A shows distributions from five different families, with the family names indicated in the legend. Panels B and C show five distributions from within the triangular and mixture families, respectively, and the legend indicates the skewness of each distribution. Panel D shows five distributions from the *t* family, and the legend indicates the number of degrees of freedom of each distribution. In panels A and D, the stimulus values have been linearly rescaled to equate the ranges of the different functions.

tional comparison of probit and SK analyses with different amounts of skew in the underlying distribution. In the present simulations, we used mixture distributions with skewness values of −0.8, −0.4, 0, 0.4, and 0.8. In fact, the mixture distribution with zero skew is simply the uniform (i.e., rectangular) distribution over the range of 0–1.

***t* distributions.** Finally, we also included five Student's *t* distributions with degrees of freedom values of 5, 6, 7, 10, and 16. These were selected because they have substantially different values of kurtosis $\gamma_2(M)$, as is shown in Table 3.

Methods of Analysis

We analyzed every simulated data set twice—once with the SK method and once with probit analysis—so that we could compare the results of two hypothetical researchers obtaining the same data but conducting different analyses. As was described earlier, for the SK analysis, we first monotonized the estimated response probabilities, if necessary, using the method of Ayer et al. (1955), and we then computed the estimated moments and percentiles of this distribution (e.g., Equation 2). The SK analysis was carried out on every simulated data set because it is a distribution-free method, in contrast to the probit analysis, which was carried out only on reasonably normal data sets, as will be described next.

The probit analysis of each simulated data set proceeded in two steps. First, the fit of the simulated data set to the normal was evaluated with a χ^2 test (Guilford, 1936). If the computed χ^2 value was significant at $p < .05$, the data set was deemed inappropriate for probit analysis and was excluded from further consideration in the evaluation of the probit estimation method.[8] It seems reasonable to exclude from the analysis any simulated data sets that were obviously nonnormal, because an alert researcher obtaining such data would realize that probit analysis would be inappropriate for them. With such data, the researcher would likely search for a transformation to achieve approximate normality, but it is beyond the scope of our simulations to attempt to model this search. We simply note that another advantage of the SK method is that it avoids this potential problem, because this method can be used regardless of the true underlying distribution.

If the simulated data set yielded a nonsignificant χ^2 value, the second step in the probit analysis was to compute the maximum likelihood estimates of the mean and standard deviation of the underlying normal distribution. For any given set of parameter values, the likelihood of the simulated data set is

$$L = \prod_{i=1}^{k} p_i^{\left(n_i - X_i\right)} \cdot \left(1 - p_i\right)^{X_i},$$

where p_i is the cumulative normal density at stimulus value s_i with the given parameter values (Finney, 1971, chap. 5). For each simulated data set, this likelihood func-

tion was maximized via Rosenbrock's (1960) numerical search algorithm.[9] The estimated mean, standard deviation, and *dl* were then computed directly from the normal distribution with the maximum likelihood parameter estimates, and the estimated median equals the estimated mean. Note that there is no point in estimating higher moments with the probit analysis, because these cannot vary with the data (e.g., estimated skewness is always 0).

For both probit analysis and the SK method, we wanted to obtain from each simulated data set not only parameter estimates, as was described above, but also estimates of the standard errors of those parameter values. Researchers analyzing single data sets use estimated standard errors to compute confidence intervals around the observed parameter values and sometimes to test hypotheses concerning true values. A complete evaluation of each type of analysis, then, should also consider the information that it provides about the standard error of each estimated parameter.

The standard errors of all parameter estimates were computed using the bootstrapping procedure (e.g., DiCiccio & Romano, 1988). In this procedure, many bootstrap replications of each data set are created by resampling a new data set from the original one with replacement. The parameter value is estimated within each bootstrap replication of the experiment, and the standard error of the parameter estimate is computed from the variability of that estimate across bootstrap replications.[10] As far as we know, the bootstrapping procedure is the only general method for computing the standard errors of parameters estimated with the SK method. Although there are alternative methods of computing the standard errors of parameters estimated with probit analysis (e.g., Finney, 1952), the bootstrapping procedure, too, performs quite well for computing those standard errors (Foster & Bischof, 1991).[11]

SIMULATION RESULTS

Location Estimators

In this section, we compared the performance of the location estimators: probit analysis's estimator of the mean $\hat{\mu}_{PR}$ and the median $\widehat{pse}_{PR}$ and the SK method's estimators of the same two parameters, $\hat{\mu}_{SK}$ and $\widehat{pse}_{SK}$.[12] Traditionally, the most common location estimator has been the median (i.e., the *pse*) estimated with probit analysis, $\widehat{pse}_{PR}$ (e.g., Guilford, 1936; Woodworth & Schlosberg, 1954), which is also identical to the mean estimated with the probit method, because of the assumed underlying normal distribution. There is, however, no clearcut theoretical or statistical justification for estimating location with the median, rather than the mean, or for computing it with probit analysis, rather than the SK method (see Sternberg & Knoll, 1973). Thus, we examined the bias and standard error of each of these four estimators, the properties of bootstrap confidence intervals computed from each, and the power of each to detect differences in parameter values between two experimental conditions.

Bias. By definition, an estimator is biased to the extent that its observed value is, on average, higher or lower than the true value. Obviously, other things being equal, an unbiased estimator is preferable to a biased one.

As an illustration of the simulation results, Table 4 shows the biases of the probit's and SK's estimates of location in the simulations with 11 stimulus levels and 300 trials. These bias values were computed as the difference between the average value of the estimator across all 30,000 simulated data sets and the true value of the parameter being estimated.[13] The results indicate that both estimators of the mean generally have extremely small biases. The one exception to this rule arises with the Naka–Rushton distribution, for which the mean estimated by probit analysis seriously underestimated the true mean. Biases were larger for the two median estimators, especially for $\widehat{pse}_{PR}$ when the underlying distribution was highly skewed.

Computations analogous to those summarized in Table 4 were carried out for all the sets of simulations defined by combinations of different numbers of stimulus levels and trials, and these are summarized in Table 5. Overall, it seems clear that the least biased estimator of location is the mean computed with the SK method. This estimator has the smallest absolute biases, averaging across distributions, for most combinations of the numbers of stimulus levels and trials.[14] Moreover, the bias for probit's mean es-

Table 4
Biases of Location Estimators in the Simulations With 11 Stimulus Levels and 300 Trials per Experiment

	Parameter and Estimator			
	μ		*pse*	
Distribution	$\hat{\mu}_{PR}$	$\hat{\mu}_{SK}$	$\widehat{pse}_{PR}$	$\widehat{pse}_{SK}$
Normal	0.000	0.000	0.000	−0.000
Quantal	**0.003**	−0.010	0.331	0.021
Naka–Rushton	−0.049	**−0.017**	0.428	0.037
Weibull	**0.000**	−0.001	0.054	0.002
Model 4	0.000	0.000	0.000	−0.018
Triangular(−0.8)	0.001	**−0.000**	−0.034	−0.004
Triangular(−0.4)	−0.000	**−0.000**	−0.006	−0.002
Triangular(0.0)	−0.000	−0.000	−0.000	−0.001
Triangular(0.4)	0.000	**−0.000**	0.005	0.001
Triangular(0.8)	−0.001	**−0.000**	0.034	0.002
Mixture(−0.8)	0.002	**−0.001**	−0.027	−0.001
Mixture(−0.4)	0.001	**−0.000**	−0.002	−0.001
Mixture(0)	0.000	0.000	0.000	−0.001
Mixture(0.4)	−0.001	**0.001**	0.002	−0.001
Mixture(0.8)	−0.002	**0.000**	0.027	−0.001
t(5)	−0.001	−0.001	−0.001	−0.001
t(6)	0.001	0.001	0.001	−0.001
t(7)	−0.000	**−0.000**	−0.000	0.000
t(10)	0.000	0.000	0.000	0.000
t(16)	0.001	0.001	0.001	**−0.001**

Note—Each bias value is the difference between the average value of the estimated parameter over 30,000 simulations and the true value of the parameter computed directly from the underlying distribution. $\hat{\mu}_{PR}$ and $\widehat{pse}_{PR}$ are estimators of the mean and median computed with probit (PR) analysis, and $\hat{\mu}_{SK}$ and $\widehat{pse}_{SK}$ are analogous estimators computed using the Spearman–Kärber method. The bias value is printed in boldface if it was smaller in absolute value than all of the other bias values in its row before rounding to the number of decimals shown.

Table 5
Summary of Absolute Biases of Location Estimators as a Function of the Number of Trials and the Number of Stimulus Levels

		Parameter and Estimator			
		μ		*pse*	
N Trials	*N* Levels	$\hat{\mu}_{PR}$	$\hat{\mu}_{SK}$	$\widehat{pse}_{PR}$	$\widehat{pse}_{SK}$
30	5	0.020	**0.013**	0.070	0.037
30	11	0.013	**0.002**	0.038	0.028
40	5	0.024	**0.012**	0.074	0.032
40	11	0.013	**0.003**	0.039	0.047
40	21	0.024	**0.003**	0.027	0.086
60	5	0.023	**0.012**	0.073	0.030
60	11	0.011	**0.002**	0.040	0.028
60	21	0.023	**0.003**	0.027	0.025
100	5	0.024	**0.013**	0.074	0.028
100	11	0.008	**0.002**	0.043	0.012
100	21	0.022	**0.003**	0.029	0.016
300	5	0.021	**0.012**	0.071	0.027
300	11	0.003	**0.002**	0.048	0.005
300	21	0.015	**0.002**	0.036	0.013
1,000	5	**0.008**	0.012	0.038	0.028
1,000	11	**0.002**	0.002	0.053	0.003
1,000	21	0.038	**0.002**	0.037	0.003

Note—The numbers in the table are the averages across the 20 simulated distributions of the absolute bias obtained with each distribution. The average is printed in boldface if it was smaller than all of the other averages in its row before rounding to the number of decimals shown. $\hat{\mu}_{PR}$ and $\widehat{pse}_{PR}$ are estimators of the mean and median computed with probit (PR) analysis, and $\hat{\mu}_{SK}$ and $\widehat{pse}_{SK}$ are analogous estimators computed using the Spearman–Kärber method.

timates was often more than twice as large as the bias obtained with the SK method. Thus, if a researcher's main goal is to obtain a minimally biased estimate of location, it appears that the best approach is to use the mean estimated with the SK method.

Standard error of measurement. By definition, the standard error of an estimator indicates its tendency to fluctuate from one random sample to the next. Other things being equal, an estimator with a smaller standard error is preferable to an estimator with a larger one.

The standard errors of the probit's and SK's location estimators were computed separately for each combination of number of stimulus levels and number of trials. Within each combination, the standard error was computed as the standard deviation of each estimator across the 30,000 simulated experiments. These standard errors are summarized in Table 6. Overall, the lowest standard errors were obtained when the mean was estimated with the SK method. Thus, a researcher interested primarily in obtaining a location estimate with low random variability would also be advised to use the mean computed with the SK method.

Confidence intervals. A researcher may want not only a point estimate of a location parameter—in which case, the estimator's bias and standard error are important—but also an interval estimate, and bootstrapping can be used to construct interval estimates. Specifically, bootstrap samples can be used to estimate the standard error of the location estimator obtained from a single observed sample, and a bootstrap 95% confidence interval can be computed as the observed value plus or minus 1.96 times the bootstrap standard error.

The accuracy of bootstrap confidence intervals can be evaluated on two criteria: (1) Confidence intervals should be as narrow as possible, and (2) they should contain the true parameter value for approximately 95% of the samples used to compute them. With respect to the first criterion, we found that the average width of the bootstrap confidence intervals computed from a given estimator were directly proportional to the standard error of that estimator (see Table 6). This is simply a reflection of the fact that the average of the bootstrap standard errors was very nearly the same as the actual standard error (i.e., the bootstrap standard error is approximately unbiased) for each estimator. Thus, $\hat{\mu}_{SK}$ yielded the narrowest confidence intervals, just as it produced the smallest standard errors of measurement in Table 6.

With respect to the second criterion, Table 7 summarizes the percentages of samples for which the true value of the location parameter was contained within the bootstrap confidence interval for it computed from each estimator. Under most conditions, confidence intervals for the mean include the true value more than 90% of the time, indicating that these confidence interval procedures do approximately capture the true value of this location parameter rather well. Confidence intervals computed with the probit method are slightly more likely to contain the true value than are those computed with the SK method, although this advantage must stem largely from the fact that probit's confidence intervals are wider owing to larger

Table 6
Summary of Standard Errors of Location Estimators as a Function of the Number of Trials and the Number of Stimulus Levels

		Parameter and Estimator			
		μ		*pse*	
N Trials	*N* Levels	$\hat{\mu}_{PR}$	$\hat{\mu}_{SK}$	$\widehat{pse}_{PR}$	$\widehat{pse}_{SK}$
30	5	0.216	**0.214**	0.216	0.277
30	11	0.211	**0.194**	0.211	0.289
40	5	0.187	**0.185**	0.187	0.246
40	11	0.183	**0.168**	0.183	0.267
40	21	0.201	**0.168**	0.201	0.284
60	5	0.151	**0.151**	0.151	0.209
60	11	0.147	**0.137**	0.147	0.232
60	21	0.161	**0.137**	0.161	0.244
100	5	0.116	0.117	0.116	0.169
100	11	0.117	**0.112**	0.117	0.197
100	21	0.122	**0.106**	0.122	0.206
300	5	0.069	**0.067**	0.069	0.104
300	11	0.066	**0.065**	0.066	0.128
300	21	0.069	**0.064**	0.069	0.148
1,000	5	0.035	0.037	0.035	0.060
1,000	11	0.031	0.035	0.031	0.081
1,000	21	0.032	0.034	0.032	0.096

Note—The numbers in the table are the averages across the 20 simulated distributions of the standard error obtained with each distribution. The average is printed in boldface if it was smaller than all of the other averages in its row before rounding to the number of decimals shown. $\hat{\mu}_{PR}$ and $\widehat{pse}_{PR}$ are estimators of the mean and median computed with probit (PR) analysis, and $\hat{\mu}_{SK}$ and $\widehat{pse}_{SK}$ are analogous estimators computed using the Spearman–Kärber method.

Table 7
Summary of Percentages of Bootstrap Confidence Intervals for Location Parameters Containing the True Parameter Value as a Function of the Number of Trials and the Number of Stimulus Levels

		Parameter and Estimator			
		μ		*pse*	
N Trials	*N* Levels	$\hat{\mu}_{PR}$	$\hat{\mu}_{SK}$	$\widehat{pse}_{PR}$	$\widehat{pse}_{SK}$
30	5	**92.4**	90.3	89.1	86.6
30	11	**88.4**	87.7	87.6	85.2
40	5	92.6	**92.7**	88.9	89.3
40	11	**90.2**	89.9	88.9	88.3
40	21	**82.9**	81.4	81.0	73.4
60	5	93.2	**93.3**	88.6	85.6
60	11	**92.5**	92.0	90.5	89.8
60	21	**89.2**	88.6	87.0	86.8
100	5	91.7	**93.9**	86.7	85.5
100	11	**93.5**	93.3	90.0	90.7
100	21	**92.4**	92.0	89.5	91.0
300	5	92.0	**93.6**	78.4	87.4
300	11	**95.0**	94.8	83.5	91.5
300	21	94.0	**94.3**	87.2	92.7
1,000	5	**92.2**	91.1	65.1	87.6
1,000	11	**95.4**	95.3	72.3	91.6
1,000	21	**95.3**	95.1	76.3	92.7

Note—The numbers in the table are the averages across the 20 simulated distributions of the percentage obtained with each distribution. The percentage is printed in boldface if it was larger than all of the other percentages in its row before rounding to the number of decimals shown. $\hat{\mu}_{PR}$ and $\widehat{pse}_{PR}$ are estimators of the mean and median computed with probit (PR) analysis, and $\hat{\mu}_{SK}$ and $\widehat{pse}_{SK}$ are analogous estimators computed using the Spearman–Kärber method.

bootstrap standard errors. Confidence intervals for the mean perform well whether estimated with probit analysis or the SK method, although the latter method appears slightly better overall. Confidence intervals for the median included the true value less often than those for the mean—substantially so when estimated with probit analysis in the simulations with 1,000 trials.

Power of comparisons of experimental conditions. Although we have so far considered only estimation of the location of a single psychometric function, in practice researchers often want to compare the locations of psychometric functions across conditions. For example, researchers might want to compare the locations of psychometric functions for duration discrimination under two different attentional conditions (Mattes & Ulrich, 1998) or of psychometric functions for weight discriminations of objects with smooth versus rough surfaces (Flanagan, Wing, Allison, & Spenceley, 1995; Rinkenauer, Mattes, & Ulrich, 1999).[15]

It is convenient to compare the powers of different estimators by using a measure analogous to the d' of signal detection theory[16] (e.g., Green & Swets, 1966). For a test using $\hat{\mu}_{PR}$, for example, this value is

$$d'\left(\hat{\mu}_{PR}\right) = \frac{E\left[\hat{\mu}_{PR,2}\right] - E\left[\hat{\mu}_{PR,1}\right]}{\sqrt{SE\left[\hat{\mu}_{PR,2}\right]^2 + SE\left[\hat{\mu}_{PR,1}\right]^2}}, \qquad (8)$$

where $E[\mu_{PR,i}]$ is the expected value of probit's mean estimator in condition i and $SE[\hat{\mu}_{PR,i}]$ is its standard error. Analogous d' values can be defined for all of the other estimators of location (i.e., $\widehat{pse}_{PR}$, $\hat{\mu}_{SK}$, $\widehat{pse}_{SK}$). For any given situation, the most sensitive estimator (i.e., the one most likely to detect a difference between conditions) would then be the one with the highest value of d'.

Because all of the location estimators are approximately unbiased, it seems clear that, in most realistic situations, the estimator with the smallest standard error would yield the largest values of d'. This analysis, then, suggests that $\hat{\mu}_{SK}$ would generally have the greatest power to detect differences among experimental conditions. Surprisingly, examination of the standard errors for individual underlying distributions indicates that $\hat{\mu}_{SK}$ sometimes has more power than $\hat{\mu}_{PR}$ even when the underlying psychometric function is normal (e.g., with 11 stimulus levels and 300 trials). Typically, one expects parametric methods to have more power than nonparametric ones when the assumptions underlying the former are met (e.g., Marascuilo & McSweeny, 1977). Evidently, however, that is not the case here. Apparently, one source of the problem with the probit analysis is that its maximum likelihood parameter estimates are only asymptotically optimal, which implies that an alternative method of analysis may be superior with samples of finite size (Finney, 1952, p. 246; Foster & Bischof, 1991). In the present simulations, the superiority of the SK method over probit analysis decreased as the number of trials increased, but the SK method was still superior even with 1,000 trials. It thus appears that the asymptotic optimality of the maximum likelihood estimators may be applicable only with larger numbers of trials than would usually be obtained in practice. As will be considered further in the General Discussion section, another source of this advantage for the SK method seems to be its assumption that the underlying distribution is bounded by s_0 and s_{k+1}, rather than unbounded, as is the normal distribution.

Summary for location. These simulations provide evidence that the best estimator of the location of a psychometric function is the mean computed with the SK method. Of the four estimators we examined, it had the smallest bias, smallest standard error of estimation, and largest d' for comparing experimental conditions. In addition, it was the most effective estimator for use in conjunction with bootstrapping, leading to the narrowest bootstrap confidence intervals and the second-highest percentage of bootstrap confidence intervals including the true value. The second-best location estimator was the mean computed with the probit method, and the median estimators were distinctly inferior on the statistical criteria we examined. Thus, these results provide statistical grounds for questioning the traditional use of $\widehat{pse}_{PR}$ as the standard estimator of the location of a psychometric function.

Dispersion Estimators

In this section, we compared the performance of four dispersion estimators: probit analysis's estimator of the

Table 8
Summary of Absolute Biases of Normalized Dispersion Estimators as a Function of the Number of Trials and the Number of Stimulus Levels

		Parameter and Estimator			
		σ		*dl*	
N Trials	*N* Levels	$\hat{\sigma}_{PR}$	$\hat{\sigma}_{SK}$	$\widehat{dl}_{PR}$	$\widehat{dl}_{SK}$
30	5	0.163	**0.036**	0.182	0.108
30	11	0.069	0.158	0.125	**0.062**
40	5	0.130	**0.034**	0.148	0.103
40	11	**0.055**	0.121	0.120	0.056
40	21	**0.068**	0.182	0.117	0.098
60	5	0.095	**0.043**	0.121	0.108
60	11	**0.041**	0.081	0.125	0.042
60	21	0.061	0.132	0.126	**0.046**
100	5	0.075	**0.055**	0.120	0.112
100	11	0.036	0.054	0.123	**0.033**
100	21	0.060	0.089	0.134	**0.049**
300	5	0.068	**0.068**	0.132	0.117
300	11	0.040	**0.018**	0.110	0.022
300	21	0.051	0.044	0.126	**0.016**
1,000	5	**0.038**	0.072	0.123	0.121
1,000	11	0.028	**0.014**	0.141	0.016
1,000	21	0.064	0.023	0.162	**0.007**

Note—The numbers in the table are the averages across the 20 simulated distributions of the absolute bias obtained with each distribution. The average is printed in boldface if it was smaller than all of the other averages in its row before rounding to the number of decimals shown. The bias values were divided by the true parameter value to equate the scales for the estimators of σ and *dl* (see note 17). $\hat{\sigma}_{PR}$ and $\widehat{dl}_{PR}$ are estimators of the standard deviation σ and difference limen *dl* computed with probit (PR) analysis, and $\hat{\sigma}_{SK}$ and $\widehat{dl}_{SK}$ are analogous estimators computed using the Spearman–Kärber method.

standard deviation $\hat{\sigma}_{PR}$ and difference limen $\widehat{dl}_{PR}$ and the SK method's estimators of the same two parameters, $\hat{\sigma}_{SK}$ and $\widehat{dl}_{SK}$. Again, we examined the bias and standard error of each estimator, its usefulness in computing bootstrap confidence intervals, and its power to detect differences in parameter values between experimental conditions.[17]

Bias. Table 8 summarizes the normalized biases of the four dispersion estimators in the different simulation conditions, but there is no clear overall best estimator in this table. Depending on the numbers of trials and stimulus levels, the least biased dispersion estimator was either $\hat{\sigma}_{PR}$, $\hat{\sigma}_{SK}$, or $\widehat{dl}_{PR}$. $\widehat{dl}_{PR}$ is the one estimator that does not perform best under any of these conditions, and it often does much worse than the others. This result is ironic because $\widehat{dl}_{PR}$ is traditionally the most common measure of dispersion. In any case, if a researcher's main goal is to obtain a minimally biased estimate of dispersion, it would seem necessary to compare the estimators via computer simulation under conditions close to those of the actual experiment.

Standard error of measurement. Table 9 summarizes the standard errors of the dispersion estimators. Averaging across distributions, $\hat{\sigma}_{SK}$ had the lowest standard error for all combinations of numbers of trials and stimuli. The relatively low standard error of this estimator strongly suggests that it should be the default dispersion estimator, especially in conjunction with the fact that this is one of the dispersion estimators with relatively low biases (see Table 8).

Confidence intervals. As was true for the location estimators, the bootstrap standard errors of the dispersion estimators were nearly unbiased estimates of the actual standard errors of these estimators, so the average normalized widths of the confidence intervals were again proportional to the standard errors of these estimators shown in Table 9. As a result, $\hat{\sigma}_{SK}$ yielded the narrowest normalized confidence intervals of the four dispersion estimators.

Table 10 summarizes the percentages of bootstrap confidence intervals containing the true value of the dispersion parameter. The results, however, are somewhat mixed. $\widehat{dl}_{SK}$ generally performed the best, but even its nominal 95% confidence intervals sometimes contained the true value quite a bit less than the expected 95% of the time. It performed particularly badly when there were 300 or 1,000 trials and only five stimulus levels. Overall, then, it appears that bootstrap confidence intervals for dispersion parameters appear to present greater risks of missing the true value, and the extent of these risks depends on the numbers of trials and stimulus levels in the experiment.

Comparison across experimental conditions. In practice, researchers also often compare the dispersions of psychometric functions across conditions, just as they do location parameters. For example, a researcher might want to

Table 9
Summary of Standard Errors of Normalized Dispersion Estimators as a Function of the Number of Trials and the Number of Stimulus Levels

		Parameter and Estimator			
		σ		*dl*	
N Trials	*N* Levels	$\hat{\sigma}_{PR}$	$\hat{\sigma}_{SK}$	$\widehat{dl}_{PR}$	$\widehat{dl}_{SK}$
30	5	0.364	**0.246**	0.380	0.383
30	11	0.364	**0.240**	0.384	0.451
40	5	0.306	**0.215**	0.322	0.352
40	11	0.305	**0.209**	0.323	0.417
40	21	0.335	**0.210**	0.354	0.428
60	5	0.234	**0.178**	0.246	0.294
60	11	0.238	**0.172**	0.254	0.355
60	21	0.259	**0.173**	0.277	0.384
100	5	0.170	**0.139**	0.179	0.227
100	11	0.188	**0.141**	0.198	0.310
100	21	0.193	**0.135**	0.207	0.323
300	5	0.093	**0.080**	0.099	0.124
300	11	0.102	**0.082**	0.105	0.200
300	21	0.110	**0.081**	0.116	0.224
1,000	5	0.050	**0.044**	0.049	0.066
1,000	11	0.050	**0.045**	0.049	0.118
1,000	21	0.052	**0.044**	0.050	0.141

Note—The numbers in the table are the averages across the 20 simulated distributions of the standard error obtained with each distribution. The average is printed in boldface if it was smaller than all of the other averages in its row before rounding to the number of decimals shown. The estimates were divided by the true parameter value to equate the scales for the estimators of σ and *dl* (see note 17). $\hat{\sigma}_{PR}$ and $\widehat{dl}_{PR}$ are estimators of the standard deviation σ and difference limen *dl* computed with probit (PR) analysis, and $\hat{\sigma}_{SK}$ and $\widehat{dl}_{SK}$ are analogous estimators computed using the Spearman–Kärber method.

Table 10
Summary of Percentages of Bootstrap Confidence Intervals for Dispersion Parameters Containing the True Parameter Value as a Function of the Number of Trials and the Number of Stimulus Levels

		Parameter and Estimator			
		σ		dl	
N Trials	N Levels	$\hat{\sigma}_{PR}$	$\hat{\sigma}_{SK}$	$\widehat{dl}_{PR}$	$\widehat{dl}_{SK}$
30	5	82.3	79.5	75.1	**87.6**
30	11	**85.0**	73.2	80.4	84.6
40	5	86.7	83.8	77.1	**90.5**
40	11	86.9	78.4	82.8	**88.0**
40	21	**82.3**	65.7	77.5	80.8
60	5	89.8	88.4	77.7	**92.7**
60	11	89.5	83.9	84.1	**91.2**
60	21	87.9	74.8	82.2	**88.1**
100	5	90.4	90.7	74.5	**93.8**
100	11	91.4	87.6	83.6	**92.9**
100	21	91.4	81.7	83.6	**92.0**
300	5	84.1	**86.9**	64.0	77.7
300	11	92.5	92.3	74.9	**94.1**
300	21	92.1	88.5	78.3	**94.6**
1,000	5	**83.0**	63.6	44.7	57.6
1,000	11	93.8	92.6	63.5	**93.8**
1,000	21	84.8	89.4	63.5	**94.8**

Note—The numbers in the table are the averages across the 20 simulated distributions of the percentage obtained with each distribution. The percentage is printed in boldface if it was larger than all of the other percentages in its row before rounding to the number of decimals shown. $\hat{\sigma}_{PR}$ and $\widehat{dl}_{PR}$ are estimators of the standard deviation σ and difference limen dl computed with probit (PR) analysis, and $\hat{\sigma}_{SK}$ and $\widehat{dl}_{SK}$ are analogous estimators computed using the Spearman–Kärber method.

compare the psychometric functions for temporal-order discriminations across two modalities (e.g., Hirsh & Sherrick, 1961) or the functions for suprathreshold load discriminations with natural teeth versus tooth implants (Mühlbradt, Mattes, Möhlmann, & Ulrich, 1994).

As was done for location estimators, we can again evaluate the power of each estimator by using d' values such as[16]

$$d'\left(\hat{\sigma}_{PR}\right) = \frac{E\left[\hat{\sigma}_{PR,2}\right] - E\left[\hat{\sigma}_{PR,1}\right]}{\sqrt{SE\left[\hat{\sigma}_{PR,2}\right]^2 + SE\left[\hat{\sigma}_{PR,1}\right]^2}}. \quad (9)$$

Although the biases of dispersion estimators are somewhat larger than those of location estimators, the numerators of these d' values can still probably be ignored, because the biases are likely to be equal across conditions and thus disappear in the subtraction. Again, then, it appears that the most powerful estimator will be the one with the smallest normalized standard error. For dispersion estimators, that is clearly $\hat{\sigma}_{SK}$.

Summary for dispersion. It appears that $\hat{\sigma}_{SK}$ is the best overall estimator of dispersion, although this generalization depends somewhat on the experimental design and goals. $\hat{\sigma}_{SK}$ is clearly best if the main goal is to detect differences in dispersion across experimental conditions because it has the smallest standard error and bootstrap standard error and probably the largest d'. On the other hand, if the experimental goal is to obtain either a point or an interval estimate of the dispersion within a given condition, the best estimator appears to depend on the numbers of stimulus levels and trials.

Skewness Estimators

In this section, we evaluated two skewness estimators provided by the SK method. No comparison with probit analysis was possible, because probit does not provide an estimator of skewness. Although skewness is rarely examined in analyses of psychometric functions (but see Ulrich, 1987), it provides information that can be useful in at least two situations. First, the parameter conveys information about the shape of the psychometric function that may have important theoretical implications. Many models predict symmetric functions (see Falmagne, 1985), and evidence of skewness would thus contradict those models. Second, when a χ^2 test (e.g., Guilford, 1936) indicates that a psychometric function deviates significantly from the normal distribution, a skewness measure may help reveal more specifically the nature of that deviation.

We considered two skewness estimators—one based on moments and the other based on percentiles—that can be estimated with the SK method. The parameters being estimated were

$$\gamma_1(M) = \frac{E\left[\left(x-\mu\right)^3\right]}{\sigma^3} \quad (10)$$

and

$$\gamma_1(P) = \frac{s_{75} - 2\cdot s_{50} + s_{25}}{s_{75} - s_{25}}. \quad (11)$$

$\gamma_1(M)$ is the third central moment of the distribution standardized by the cube of the standard deviation, and $\gamma_1(P)$ is a standard quartile-based measure of skewness (e.g., Spiegel, 1975), where s_p is the pth percentile of the distribution. As with location and dispersion parameters, we examined the bias and standard error of these estimators.[18]

Bias. Table 11 summarizes the absolute biases of the two skewness estimators as a function of the number of stimulus levels and trials. These biases were clearly much larger than those obtained with location and dispersion estimators, and this may be due to the fact that it is difficult to recover exact shape information from a limited set of stimulus values. Interestingly, the biases in $\hat{\gamma}_1(M)_{SK}$ did not depend much on the number of stimulus levels or—at least within the range of 60–1,000—trials. In contrast, the biases in $\gamma_1(P)$ appear to decrease with increases in the numbers of trials and stimulus levels.

Although it cannot be seen in the table, inspection of skewness estimates for each distribution separately indicated that these estimators—especially $\hat{\gamma}_1(M)_{SK}$—tend to attenuate the true skewness (i.e., underestimate positive skewness values and overestimate negative ones). Nonetheless, the mean estimated skewnesses had the correct sign for each skewed distribution for nearly all combinations of numbers of stimulus levels and trials, indicating that the skewness estimators can at least be used to recover

Table 11
Summary of Absolute Biases of Skewness Estimators as a Function of the Number of Trials and the Number of Stimulus Levels

		Parameter and Estimator	
N Trials	*N* Levels	$\hat{\gamma}_1(M)_{SK}$	$\hat{\gamma}_1(P)_{SK}$
30	5	0.266	0.039
30	11	0.253	0.047
40	5	0.237	0.040
40	11	0.226	0.043
40	21	0.219	0.119
60	5	0.204	0.037
60	11	0.196	0.035
60	21	0.196	0.041
100	5	0.193	0.036
100	11	0.197	0.024
100	21	0.200	0.029
300	5	0.201	0.036
300	11	0.201	0.011
300	21	0.203	0.017
1,000	5	0.207	0.039
1,000	11	0.204	0.008
1,000	21	0.206	0.007

Note—The numbers in the table are the averages across the 20 simulated distributions of the absolute bias obtained with each distribution. $\hat{\gamma}_1(M)_{SK}$ and $\hat{\gamma}_1(P)_{SK}$ are the moment-based and percentile-based estimators of skewness computed using the Spearman–Kärber method.

qualitative information about skewness. Furthermore, the values of $\hat{\gamma}_1(M)_{SK}$ obtained in individual samples had the correct sign in more than half of the samples for virtually all combinations of numbers of trials and stimulus levels, suggesting that this measure would have some power to detect skewness in multiparticipant tests of symmetry (e.g., a *t* or sign test computed across participants).

Standard error of measurement. Table 12 summarizes the standard errors of the skewness estimators. Again, these standard errors are considerably larger than those of the location or dispersion estimators, and this is unsurprising because estimates of higher moments and extreme percentiles tend to be unreliable (e.g., Stuart & Ord, 1987, p. 338). The standard error of $\hat{\gamma}_1(M)_{SK}$ decreased with the number of trials, as was expected, but was virtually unaffected by the number of stimulus levels. The standard error of $\hat{\gamma}_1(P)_{SK}$ also decreased with the number of trials and, in addition, increased with the number of stimulus levels.

Summary for skewness. The SK method's two skewness estimators were biased toward zero, so estimated skewness values tend to be attenuated relative to true skewness values. Nonetheless, these skewness estimators can be used to assess deviations from symmetry, because they tend to have the correct sign.

Kurtosis Estimators

In this section we evaluated two of the SK method's kurtosis estimators. Like skewness, the kurtosis parameter is of interest as a descriptor of the shape of the psychometric function, especially when a nonnormal distribution is predicted theoretically or when the normal distribution is rejected by a χ^2 test. Again, no comparison with probit analysis was possible, because that type of analysis does not provide such estimators.

We considered estimators of two kurtosis parameters—one based on moments and the other based on percentiles. These parameters are

$$\gamma_2(M) = \frac{E\left[(x-\mu)^4\right]}{\sigma^4} \tag{12}$$

and

$$\gamma_2(P) = \frac{dl}{s_{90} - s_{10}}. \tag{13}$$

$\gamma_2(M)$ is the fourth central moment of the distribution standardized by the square of the variance (see DeCarlo, 1997). Note that some texts adjust this definition by subtracting three so that the normal distribution has an adjusted kurtosis of zero, but we have not adopted this convention. $\gamma_2(P)$ is a standard percentile-based measure of kurtosis (e.g., Spiegel, 1975).

Bias. Average absolute biases obtained for the two kurtosis estimators are summarized in Table 13. Perhaps unsurprisingly, these biases are quite large. Moreover, checks of the results for individual distributions indicated that when the true kurtosis value was large, $\hat{\gamma}_1(M)_{SK}$ seriously underestimated it under all simulation conditions. The degree of underestimation decreased only slightly with the number of trials, and not at all with the number of stimulus levels. Thus, it would appear that experimenters can hope to estimate accurate numerical values of $\gamma_2(P)$, but not $\gamma_2(M)$.

Standard error of measurement. Table 14 summarizes the standard errors of the two kurtosis estimators. The standard errors of $\hat{\gamma}_2(M)_{SK}$ are quite large, even larger than that of the moment-based skewness estimator, whereas the standard errors of the percentile-based estimator

Table 12
Summary of Standard Errors of Skewness Estimators as a Function of the Number of Trials and the Number of Stimulus Levels

		Parameter and Estimator	
N Trials	*N* Levels	$\hat{\gamma}_1(M)_{SK}$	$\hat{\gamma}_1(P)_{SK}$
30	5	0.620	0.283
30	11	0.723	0.485
40	5	0.586	0.265
40	11	0.665	0.480
40	21	0.713	0.584
60	5	0.519	0.241
60	11	0.570	0.444
60	21	0.599	0.538
100	5	0.429	0.207
100	11	0.481	0.403
100	21	0.473	0.496
300	5	0.262	0.134
300	11	0.289	0.303
300	21	0.289	0.383
1,000	5	0.145	0.078
1,000	11	0.161	0.199
1,000	21	0.159	0.250

Note—The numbers in the table are the averages across the 20 simulated distributions of the standard error obtained with each distribution. $\hat{\gamma}_1(M)_{SK}$ and $\hat{\gamma}_1(P)_{SK}$ are the moment-based and percentile-based estimators of skewness computed using the Spearman–Kärber method.

Table 13
Summary of Absolute Biases of Kurtosis Estimators as a Function of the Number of Trials and the Number of Stimulus Levels

		Parameter and Estimator	
N Trials	*N* Levels	$\hat{\gamma}_1(M)_{SK}$	$\hat{\gamma}_1(P)_{SK}$
30	5	2.148	0.016
30	11	2.185	0.031
40	5	2.037	0.013
40	11	2.054	0.014
40	21	2.064	0.030
60	5	1.910	0.017
60	11	1.908	0.012
60	21	1.942	0.025
100	5	1.765	0.018
100	11	1.788	0.006
100	21	1.820	0.009
300	5	1.545	0.018
300	11	1.568	0.005
300	21	1.636	0.007
1,000	5	1.454	0.018
1,000	11	1.447	0.003
1,000	21	1.492	0.003

Note. The numbers in the table are the averages across the 20 simulated distributions of the absolute bias obtained with each distribution. $\hat{\gamma}_1(M)_{SK}$ and $\hat{\gamma}_1(P)_{SK}$ are the moment-based and percentile-based estimators of kurtosis computed using the Spearman–Kärber method.

$\hat{\gamma}_2(P)_{SK}$ are reasonably small, even smaller than those of the percentile-based skewness estimator. On the other hand, it may be inappropriate to conclude that the moment-based estimator is more variable than the percentile-based estimator, because these two estimators are on rather different scales. For example, the standard errors of $\hat{\gamma}_2(M)_{SK}$ are generally a smaller proportion of the true value than are the standard errors of $\hat{\gamma}_2(P)_{SK}$.

Overall, it seems clear that increasing the number of trials decreases the standard error of the moment-based estimator more than that of the percentile-based estimator. Moreover, the standard error of the moment-based estimator appears insensitive to the number of stimulus levels, whereas the standard error of the percentile-based estimator was clearly smallest with only five levels.

Summary for kurtosis. The results indicate that it is extremely difficult to get an accurate moment-based estimate of kurtosis from psychometric functions obtained in experiments with the numbers of trials and stimulus levels considered here. In situations in which such information is required, then, it will apparently be necessary to employ a much larger experiment. In particular, informal simulations indicate that many more stimulus levels—possibly as many as 100—are needed to get unbiased moment-based kurtosis estimates. The percentile-based kurtosis estimates, on the other hand, are considerably less biased and less variable. Naturally, this suggests that theoretical analyses should, whenever possible, focus on these as the preferred measure of kurtosis.

GENERAL DISCUSSION

With the simulations reported in this article, we examined the statistical properties of a number of summary measures of observed psychometric functions. We examined summary measures of location, dispersion, skewness, and kurtosis, and for each type of measure we examined the performance of estimators of both moment-based and percentile-based parameters (e.g., mean and median as location parameters). For measures of location and dispersion, we compared the performance of estimators computed using two alternative techniques: probit analysis and the SK method. For measures of skewness and kurtosis, we evaluated the performance of estimators computed using only the SK method, because probit analysis does not provide estimators of these measures. The simulation results suggest a number of changes to standard practice in the analysis of psychometric functions.

The overall conclusion suggested by our results is that researchers should not rely exclusively on probit analysis, the current standard procedure for analysis of psychometric function data, but instead should often consider using the SK method in addition to or instead of probit analysis. The SK method has previously been used only to overcome limitations of probit analysis (e.g., to obtain independent estimates of mean and median or estimates of higher moments; Ulrich, 1987). The present results indicate, however, that estimators computed using this method often have better statistical properties than those computed using probit analysis, sometimes even when the assumptions underlying probit analysis are met. Thus, these estimators may be quite useful in routine analysis, not just when probit analysis is incapable of providing the desired estimates. Of course, probit analyses could still be conducted and reported as well, especially to facilitate comparisons with previous studies.

Table 15 summarizes the results for estimators of location and dispersion. For each type of parameter, this table

Table 14
Summary of Standard Errors of Kurtosis Estimators as a Function of the Number of Trials and the Number of Stimulus Levels

		Parameter and Estimator	
N Trials	*N* Levels	$\hat{\gamma}_1(M)_{SK}$	$\hat{\gamma}_1(P)_{SK}$
30	5	0.965	0.065
30	11	1.270	0.108
40	5	0.999	0.065
40	11	1.244	0.103
40	21	1.399	0.122
60	5	0.985	0.055
60	11	1.131	0.090
60	21	1.232	0.107
100	5	0.906	0.044
100	11	1.004	0.081
100	21	1.006	0.093
300	5	0.634	0.026
300	11	0.664	0.056
300	21	0.646	0.067
1,000	5	0.358	0.014
1,000	11	0.385	0.035
1,000	21	0.373	0.044

Note—The numbers in the table are the averages across the 20 simulated distributions of the standard error obtained with each distribution. $\hat{\gamma}_1(M)_{SK}$ and $\hat{\gamma}_1(P)_{SK}$ are the moment-based and percentile-based estimators of kurtosis computed using the Spearman–Kärber method.

shows the estimator that performed best with respect to each of four common experimental objectives. It is striking that the SK estimators performed best in most cells of the table. These include most of the common experimental goals of (1) obtaining unbiased and low-variability point estimates of the threshold and slope of a psychometric function and (2) comparing these values across two experimental conditions. Clearly, then, the superior statistical properties of the SK estimators (e.g., small bias and standard error) indicate that they deserve more widespread use in the analysis of such data.

The statistical properties of the SK method's skewness and kurtosis estimators have not previously been examined, although at least one has been previously used to summarize observed psychometric functions (e.g., Ulrich, 1987). Fortunately, our simulation results indicate that these estimators also have reasonably good statistical properties. Not surprisingly, these estimators were more biased and variable than the estimators of location and dispersion. Nonetheless, the estimators did provide at least some qualitative information that could be useful in characterizing psychometric functions—especially their deviations from normality—and in testing psychophysical models.

Limitations and Caveats

Of course, the conclusions of any simulation study are limited to some extent by the nature of the simulation conditions examined. Given the fair consistency of the SK method's superiority across the different numbers of stimulus levels and experimental trials, it seems fairly safe to suggest that our conclusions will generalize to other values of these variables, at least within the ranges studied here. Similarly, given the consistency across psychometric functions, it seems safe to expect that the conclusions will also be fairly independent of the true psychometric function.

On the other hand, the present results may depend somewhat on the precise stimuli used to sample the psychometric function. Simulations using the adaptive procedure described by Kaernbach (2001), for example, indicate that there are substantial biases in variability (slope) estimates obtained with both probit analysis and the SK method. Furthermore, even using the method of constant stimuli, one important variable that we have not yet addressed involves the exact positions of the stimulus values tested in the experiment. For the simulations reported earlier, we assumed that the stimulus values covered almost the entire transition zone of the psychometric function and that these values were located somewhat symmetrically (i.e., the smallest and largest stimulus values were at the 1% and 99% points of the function, and the middle stimulus value was at the 50% point for symmetric distributions). It is likely, however, that the SK method's estimators are sensitive to the exact stimulus placements; in particular, estimators of variance and higher moments could be seriously biased if the range of stimulus values does not adequately cover the transition zone (i.e., if there is truncation error). Fortunately, this problem is relatively easy to avoid by adequate preliminary testing of extreme stimulus levels. Nonetheless, it seemed important to determine how sensitive the SK method is to deviations from these assumptions. To this end, we conducted additional simulations with the most extreme stimulus values located at the following combinations of low and high percentile values: (2,99), (4,99), (6,99), (8,99), (1,98), (1,96), (1,94), and (1,92). In all cases, the stimulus values were equally spaced in Z scores, so the middle stimulus value also tended to differ from the 50% point in these simulations. For each pair of extreme stimulus locations, we simulated 300-trial experiments with all 20 underlying distributions and with 5, 11, and 21 stimulus levels.

Table 16 shows the results of these additional simulations for the underlying normal distribution. For location estimators, the effects of these stimulus asymmetries were systematic but tiny. For the dispersion, skewness, and kurtosis estimators, the effects were larger, as was expected. Nonetheless, small asymmetries seem to introduce only tolerable bias, suggesting that the method can be useful with only a reasonable amount of pretesting of stimulus levels. As a rule of thumb, we tentatively suggest that the smallest and largest stimulus values should be chosen to ensure $p \leq .02$ and $p \geq .98$, respectively.

Another limitation of our conclusions is that we considered the moment- and percentile-based estimators together when trying to determine the single best location or dispersion estimator. In some situations, however, testing of specific theories might require the use of a percentile-based estimator, and it would not be appropriate to consider moment-based estimators. In fact, when attention was restricted to the percentile-based estimators, probit analysis sometimes outperformed the SK method. For example, both $\widehat{pse}_{PR}$ and $\widehat{dl}_{PR}$ had smaller standard errors than did $\widehat{pse}_{SK}$ and $\widehat{dl}_{SK}$, respectively. This suggests that probit analysis could be the more useful technique in situations that require low-variability percentile-based estimators. On the other hand, percentile-based estimators are most likely to be of interest in situations in which the underlying assumption of normality is suspect, and researchers should tend to avoid probit analysis in these situations. As will be discussed further in the next section, probit's percentile-based estimators are inherently rather biased when the underlying distribution is asymmetric.

Table 15
Preferred Estimator as a Function of the Type of Parameter Being Estimated and the Primary Experimental Objective

	Primary Experimental Objective			
Type of Parameter	Minimal Bias	Minimal *SE*	Accurate Confidence Intervals	Maximal Power
Location	$\hat{\mu}_{SK}$	$\hat{\mu}_{SK}$	$\hat{\mu}_{SK}$	$\hat{\mu}_{SK}$
Dispersion	?	$\hat{\sigma}_{SK}$	$\widehat{dl}_{SK}$	$\hat{\sigma}_{SK}$

Note—$\hat{\mu}_{SK}$ and $\hat{\sigma}_{SK}$ are the estimators of μ and σ computed using the Spearman–Kärber method. ? indicates that no single method was clearly best overall.

Table 16
Effects of Asymmetric Stimulus Locations on the Spearman–Kärber Method's Estimates of Distributional Parameters as a Function of the Number of Stimulus Levels

Parameter	True Value	*N* of Levels	Percentile Values of Smallest and Largest Stimuli								
			1,99	2,99	4,99	6,99	8,99	1,92	1,94	1,96	1,98
μ	0.500	5	0.500	0.500	0.501	0.501	0.502	0.498	0.499	0.499	0.500
		11	0.500	0.500	0.502	0.503	0.505	0.495	0.497	0.498	0.499
		21	0.500	0.501	0.502	0.505	0.507	0.493	0.496	0.497	0.499
pse	0.500	5	0.500	0.499	0.500	0.500	0.500	0.500	0.500	0.500	0.501
		11	0.499	0.500	0.499	0.499	0.499	0.499	0.499	0.500	0.499
		21	0.496	0.496	0.495	0.495	0.495	0.495	0.495	0.495	0.495
σ	0.250	5	0.274	0.271	0.267	0.263	0.260	0.260	0.263	0.267	0.271
		11	0.248	0.247	0.243	0.240	0.236	0.237	0.240	0.243	0.247
		21	0.241	0.239	0.235	0.232	0.228	0.228	0.232	0.235	0.239
dl	0.169	5	0.192	0.187	0.182	0.180	0.179	0.179	0.180	0.182	0.188
		11	0.171	0.171	0.170	0.170	0.170	0.170	0.170	0.170	0.171
		21	0.167	0.167	0.167	0.167	0.166	0.167	0.166	0.167	0.167
$\gamma_1(M)$	0.000	5	0.001	0.006	0.035	0.066	0.101	−0.101	−0.067	−0.037	−0.009
		11	0.001	0.025	0.078	0.134	0.185	−0.184	−0.132	−0.079	−0.024
		21	−0.001	0.032	0.093	0.154	0.213	−0.210	−0.154	−0.093	−0.034
$\gamma_1(P)$	0.000	5	0.001	0.028	0.036	0.022	0.001	−0.001	−0.020	−0.037	−0.027
		11	0.006	−0.001	0.009	0.004	0.003	0.004	0.003	−0.001	0.006
		21	0.007	0.012	0.010	0.012	0.013	0.010	0.012	0.010	0.012
$\gamma_2(M)$	3.000	5	2.913	2.878	2.817	2.756	2.712	2.708	2.759	2.817	2.876
		11	2.720	2.660	2.584	2.517	2.476	2.474	2.519	2.576	2.664
		21	2.605	2.547	2.467	2.410	2.383	2.380	2.414	2.466	2.548
$\gamma_2(P)$	0.263	5	0.281	0.272	0.264	0.263	0.265	0.266	0.263	0.263	0.272
		11	0.267	0.267	0.268	0.269	0.270	0.270	0.268	0.267	0.268
		21	0.270	0.270	0.270	0.271	0.273	0.274	0.271	0.270	0.270

Note—The distributional parameters are the mean μ, median *pse*, standard deviation σ, difference limen *dl*, moment-based measure of skewness $\gamma_1(M)$, percentile-based measure of skewness $\gamma_1(P)$, moment-based measure of kurtosis $\gamma_2(M)$, and percentile-based measure of kurtosis $\gamma_2(P)$. The numbers in the table are the mean values of the Spearman–Kärber method's estimator of the parameter across 30,000 simulated experiments with an underlying normal distribution and the indicated number of stimulus levels and percentile values of the most extreme stimuli.

Clearly, if a situation requires unbiased percentile-based estimators with an asymmetric underlying function, the SK method would still provide the superior estimators.

Performance of Probit Analysis

Given the ubiquity of probit analysis in research using psychometric functions, an important question is whether the present simulation results provide any cause for concern about previous estimates obtained from this method of analysis. With respect to moment-based estimators, the results raise no concerns. Probit's estimators of μ and σ performed nearly as well as those computed from the SK method even when the assumption of normality was seriously violated. The relatively unbiased estimation of means was also previously reported by Simpson (1995), but the small bias in estimates of standard deviation does not seem to have been noted previously.

With respect to the percentile-based estimators $\widehat{pse}_{PR}$ and $\widehat{dl}_{PR}$, however, the results do raise a problem that might have contaminated some previous studies. Specifically, when the underlying distribution is asymmetric, both of these estimators can be seriously biased. Naturally, the biases result from probit analysis's assumption of a symmetric underlying distribution (i.e., the normal). This assumption constrains the estimate of the median to equal that of the mean, and the estimate of the median is inevitably biased toward the true mean as a result. Similarly, this assumption constrains the relation between $\widehat{pse}_{PR}$ and $\widehat{dl}_{PR}$, biasing the former toward a fixed multiple of the latter. Whatever the source of the biases, however, their existence would call into question or invalidate conclusions based on the precise numerical values of the percentile-based estimators. Because these biases are likely to be approximately the same in all experimental conditions, however, they would be unlikely to raise serious problems for conclusions about the effects of experimental manipulations.

Another issue of importance concerning probit analysis is its performance in situations satisfying the assumption of an underlying normal distribution. Somewhat surprisingly, probit analysis need not outperform the SK method in such situations, as is illustrated by Tables 17 and 18. The values in these two tables were computed in exactly the same way as the values in Tables 5, 6, 8, and 9, except that only the simulations with the underlying normal distribution were included—the results were not averaged across the different underlying psychometric functions. It is striking that the standard errors of estimates of μ and σ are consistently smaller when estimates are computed with the SK method than when they are computed with probit analysis. We believe that this happens mainly because the SK method assumes that the underlying distribution is bounded, whereas the probit method does not. This conjecture is supported by the results of simulations

varying the locations of the most extreme stimuli, s_1 and s_k, and the bounding values, s_0 and s_{k+1} (see Equation 2). When more extreme stimulus values were used, the standard errors of estimates obtained with the SK method increased more than those obtained with probit analysis, reducing or even reversing the advantage for the SK method seen in Tables 17 and 18.

Application of the Spearman–Kärber Method to Two-Alternative Forced-Choice Tasks

Psychophysicists sometimes prefer the two-alternative forced-choice task (2AFC) over the classical yes–no task for determining detection thresholds or *dl*s (Green & Swets, 1966), because the 2AFC design eliminates certain response biases that may contaminate the determination of these values in the yes–no task (e.g., the tendency to say "yes" even when no stimulus is presented). Psychometric functions obtained in the 2AFC task are not equivalent to those obtained in yes–no tasks, however, so they cannot be analyzed directly with the SK method. Fortunately, psychometric functions obtained in the 2AFC task can easily be transformed into a form appropriate for analysis with the SK method. In this section, we describe the appropriate transformation for use with 2AFC tasks examining both detection and discrimination performance. The straightforward extension to tasks with more than two alternatives is also noted briefly in note 19.

In 2AFC detection and discrimination tasks, each trial is composed of two well-defined time intervals. In the detection task, only one stimulus is presented, and it can appear in the first or the second interval with equal probabilities. The observer is asked to report the interval in which the stimulus was presented. If stimulus intensity is so low that the stimulus is never detected, the subject must guess, and thus the probability of a correct response is .5. The probability of a correct response increases from .5 to 1.0 as stimulus intensity is increased, and the detection threshold is usually defined as the stimulus intensity at which the stimulus is correctly located in 75% of all trials (e.g., Purcell & Stewart, 1988).

In a 2AFC discrimination task, two stimuli differing on a certain physical dimension are presented one at a time in the two successive intervals, with the order of presentation varying randomly from trial to trial. The observer is instructed to report the interval in which the more extreme stimulus occurred. For example, the two stimuli may differ in intensity, and the observer may be asked to report which interval contained the more intense stimulus. A score of 50% correct (chance performance) indicates a complete lack of discriminability, and the physical stimulus difference that is just sufficient to produce 75% correct responses is usually defined as the *dl* (e.g., Barrett, Whitaker, McGraw, & Herbert, 1999).

With respect to the SK method, the problem with the 2AFC task is that it produces psychometric functions that range from .5 to 1 and that are, therefore, not strictly comparable with the classical 0–1 psychometric functions considered to this point. Fortunately, it is relatively easy to adapt the psychometric function obtained in the 2AFC task for use with the SK method. The psychometric function

Table 17
Biases and Standard Errors of Location Estimators in Simulations With an Underlying Normal Psychometric Function as a Function of the Number of Trials and the Number of Stimulus Levels

		Biases (Parameter and Estimator)				Standard Errors (Parameter and Estimator)			
		μ		*pse*		μ		*pse*	
N Trials	*N* Levels	$\hat{\mu}_{PR}$	$\hat{\mu}_{SK}$	$\widehat{pse}_{PR}$	$\widehat{pse}_{SK}$	$\hat{\mu}_{PR}$	$\hat{\mu}_{SK}$	$\widehat{pse}_{PR}$	$\widehat{pse}_{SK}$
30	5	0.000	0.000	0.000	0.003	0.083	0.083	0.083	0.105
30	11	0.001	**0.001**	0.001	0.011	0.076	**0.074**	0.076	0.111
40	5	0.000	0.000	0.000	0.001	0.070	0.071	0.070	0.091
40	11	0.000	**0.000**	0.000	0.019	0.065	**0.064**	0.065	0.103
40	21	0.000	**0.000**	0.000	0.037	0.065	**0.064**	0.065	0.109
60	5	0.001	0.000	0.001	**0.000**	0.058	0.058	0.058	0.076
60	11	0.000	**0.000**	0.000	0.011	0.053	**0.052**	0.053	0.088
60	21	0.000	**0.000**	0.000	0.010	0.053	**0.052**	0.053	0.092
100	5	0.000	**0.000**	0.000	0.000	0.045	0.046	0.045	0.063
100	11	0.000	0.000	0.000	0.003	0.043	**0.043**	0.043	0.073
100	21	0.000	0.000	0.000	0.006	0.041	**0.041**	0.041	0.077
300	5	0.000	0.000	0.000	**0.000**	0.026	0.026	0.026	0.040
300	11	0.000	0.000	0.000	0.000	0.025	**0.025**	0.025	0.045
300	21	0.000	**0.000**	0.000	0.005	0.024	**0.024**	0.024	0.054
1,000	5	0.000	**0.000**	0.000	0.000	0.014	0.014	0.014	0.024
1,000	11	0.000	**0.000**	0.000	0.000	0.013	**0.013**	0.013	0.026
1,000	21	0.000	**0.000**	0.000	0.001	0.013	**0.013**	0.013	0.033

Note—Each bias value is the difference between the average value of the estimated parameter over 30,000 simulations and the true value of the parameter computed directly from the underlying distribution. Each standard error is the standard deviation of the indicated estimator over 30,000 simulations. $\hat{\mu}_{PR}$ and $\widehat{pse}_{PR}$ are estimators of the mean and median computed with probit (PR) analysis, and $\hat{\mu}_{SK}$ and $\widehat{pse}_{SK}$ are analogous estimators computed using the Spearman–Kärber method. The bias or standard error value is printed in boldface if it was smaller in absolute value than all of the other bias or standard error values in its row before rounding to the number of decimals shown.

Table 18
Biases and Standard Errors of Dispersion Estimators in Simulations With an Underlying Normal Psychometric Function as a Function of the the Number of Trials and the Number of Stimulus Levels

		Biases (Parameter and Estimator)				Standard Errors (Parameter and Estimator)			
		σ		*dl*		σ		*dl*	
N Trials	*N* Levels	$\hat{\sigma}_{PR}$	$\hat{\sigma}_{SK}$	$\widehat{dl}_{PR}$	$\widehat{dl}_{SK}$	$\hat{\sigma}_{PR}$	$\hat{\sigma}_{SK}$	$\widehat{dl}_{PR}$	$\widehat{dl}_{SK}$
30	5	0.155	**0.006**	0.155	0.088	0.396	**0.267**	0.396	0.408
30	11	0.091	0.155	0.090	**0.018**	0.327	**0.254**	0.327	0.472
40	5	0.109	**0.035**	0.109	0.096	0.332	**0.234**	0.332	0.379
40	11	0.065	0.114	0.064	**0.034**	0.273	**0.222**	0.273	0.443
40	21	0.059	0.183	**0.058**	0.091	0.282	**0.223**	0.282	0.452
60	5	0.066	**0.058**	0.066	0.116	0.251	**0.193**	0.251	0.308
60	11	0.038	0.069	0.037	**0.013**	0.214	**0.182**	0.214	0.375
60	21	0.036	0.129	0.036	**0.034**	0.220	**0.183**	0.220	0.404
100	5	0.033	0.080	**0.033**	0.132	0.176	**0.149**	0.176	0.225
100	11	0.027	0.044	0.027	**0.013**	0.169	**0.148**	0.169	0.329
100	21	0.021	0.085	**0.021**	0.045	0.165	**0.142**	0.165	0.339
300	5	0.011	0.097	**0.011**	0.140	0.097	**0.086**	0.097	0.104
300	11	0.008	**0.006**	0.008	0.014	0.096	**0.087**	0.096	0.214
300	21	0.008	0.038	**0.008**	0.009	0.096	**0.085**	0.096	0.234
1,000	5	0.004	0.102	**0.004**	0.141	0.053	**0.047**	0.053	0.051
1,000	11	0.002	0.006	**0.002**	0.017	0.052	**0.047**	0.052	0.130
1,000	21	0.002	0.016	0.002	**0.002**	0.051	**0.046**	0.051	0.147

Note—Each bias value is the difference between the average value of the estimated parameter over 30,000 simulations and the true value of the parameter computed directly from the underlying distribution. Each standard error is the standard deviation of the indicated estimator over 30,000 simulations. $\hat{\sigma}_{PR}$ and $\widehat{dl}_{PR}$ are estimators of the standard deviation σ and difference limen *dl* computed with probit (PR) analysis, and $\hat{\sigma}_{SK}$ and $\widehat{dl}_{SK}$ are analogous estimators computed using the Spearman–Kärber method. The bias or standard error value is printed in boldface if it was smaller in absolute value than all of the other bias or standard error values in its row before rounding to the number of decimals shown.

$G(s)$ of a 2AFC task can be represented in terms of a CDF $F(s)$ (e.g., Harvey, 1986):

$$G(s) = .5 + .5 \cdot F(s), \tag{14}$$

where $G(s)$ increases from .5 to 1 as $F(s)$ increases from 0 to 1. Note that the 75% threshold associated with the function $G(s)$ corresponds to the median of $F(s)$ because .75 = .5 + .5 · .5. Furthermore, the above expression can be rearranged to obtain

$$F(s) = 2 \cdot G(s) - 1. \tag{15}$$

Thus, an observed set of correct response probabilities in a 2AFC task, $\hat{g}_i, i = 1, \ldots, k$, can be transformed to the corresponding probability estimates needed for the SK method, using the transformation

$$\hat{p}_i = 2 \cdot \hat{g}_i - 1, \tag{16}$$

and the SK method can then be used to estimate any desired parameter from these $\hat{p}_i$ values.[19] As an illustration, assume that a 2AFC task yielded correct response probabilities of $\hat{g}$ = (.5, .515, .575, .775, .935, 1.0) at the stimulus levels of s = (3, 6, 9, 12, 15, 18). The transformation yields $\hat{p}$ = (0, .03, .15, .55, .87, 1.0), and application of the standard SK method with these $\hat{p}$ values yields an estimated mean and median of 11.7 and 11.6, respectively.

In sum, then, the above analysis suggests that the SK method may also be applicable to 2AFC tasks and, in general, *m*-AFC tasks. It is not yet certain, however, whether the SK method would also generate stable parameter estimates with such tasks. Because the present simulations were not modeled to capture this exact situation, our results can only encourage this conjecture. A possible complication with the extension of the SK method to forced-choice tasks, however, is that the method might benefit from differential weightings of the response probabilities at each stimulus level.[20] In particular, response probabilities closer to the guessing level might need to be given smaller weights in order to attenuate random variations owing to guessing in the lower part of the psychometric function and, thus, to increase the stability of the estimates computed from this function. Clearly, further work is needed to evaluate this extension of the SK method.

CONCLUSIONS

The SK method is a distribution-free method for the analysis of psychometric functions. This fact implies that it can be confidently applied when the shape of the underlying psychometric function is either unknown or known to be nonnormal. It also enables the method to provide a very wide range of estimates, including estimates of skewness and higher moments, separate estimates of the mean and median, and independent estimates of σ and *dl*. In contrast, probit analysis, which is the standard technique for the analysis of psychometric functions, is a parametric technique that shares neither of these two features. In addition, the SK method's estimators tend to have excellent statistical properties, especially low bias and standard error. It seems clear, then, that the SK method has at-

tractive features that make it worthy of further investigation and use as a tool in the analysis of psychometric functions.

REFERENCES

Allan, L. G. (1975). Temporal order psychometric functions based on confidence-rating data. *Perception & Psychophysics*, **18**, 369-372.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., & Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics*, **26**, 641-647.

Barrett, B. T., Whitaker, D., McGraw, P. V., & Herbert, A. M. (1999). Discriminating mirror symmetry in foveal and extra-foveal vision. *Vision Research*, **39**, 3737-3744.

Church, J. D., & Cobb, E. B. (1973). On the equivalence of the Spearman–Kärber and maximum likelihood estimates of the mean. *Journal of the American Statistical Association*, **68**, 201-202.

Cornell, R. G. (1983). Kärber method. In S. Kotz, N. L. Johnson, & C. B. Read (Eds.), *Encyclopedia of statistical sciences* (Vol. 4, pp. 354-357). New York: Wiley.

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods*, **2**, 292-307.

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

Epstein, B., & Churchman, C. W. (1944). On the statistics of sensitivity data. *Annals of Mathematical Statistics*, **15**, 90-96.

Falmagne, J. C. (1985). *Elements of psychophysical theory*. Oxford: Oxford University Press.

Fechner, G. T. (1860). *Elemente der Psychophysik* [Elements of psychophysics]. Leipzig: Breitkopf und Härtel.

Finney, D. J. (1952). *Probit analysis: A statistical treatment of the sigmoid response curve* (2nd ed.). Cambridge: Cambridge University Press.

Finney, D. J. (1971). *Probit analysis: A statistical treatment of the sigmoid response curve* (3rd ed.) Cambridge: Cambridge University Press.

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

Flanagan, J. R., & Wing, A. M. (1997). Effects of surface texture and grip force on the discrimination of hand-held loads. *Perception & Psychophysics*, **59**, 111-118.

Flanagan, J. R., Wing, A. M., Allison, S., & Spencely, A. (1995). Effects of surface texture on weight perception when lifting objects with a precision grip. *Perception & Psychophysics*, **57**, 282-290.

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

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.

Gescheider, G. A. (1997). *Psychophysics: The fundamentals* (3rd ed.). Hillsdale, NJ: Erlbaum.

Getty, D. J. (1975). Discrimination of short temporal intervals: A comparision of two models. *Perception & Psychophysics*, **18**, 1-8.

Green, D. M., & Luce, R. D. (1975). Parallel psychometric functions from a set of independent detectors. *Psychological Bulletin*, **82**, 483-486.

Green, D. M., & McGill, W. J. (1970). On the equivalence of detection probabilities and well-known statistical quantities. *Psychological Review*, **77**, 294-301.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Guilford, J. P. (1936). *Psychometric methods*. New York: McGraw-Hill.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

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

Hirsh, I. J., & Sherrick, C. E., Jr. (1961). Perceived order in different sense modalities. *Journal of Experimental Psychology*, **62**, 423-432.

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

Kärber, G. (1931). Beitrag zur kollektiven Behandlung pharmakologischer Reihenversuche [A contribution to the collective treatment of a pharmacological experimental series]. *Archiv für experimentelle Pathologie und Pharmakologie*, **162**, 480-483.

Killeen, P. R., Fetterman, J. G., & Bizo, L. A. (1997). Time's causes. In C. M. Bradshaw & E. Szabadi (Eds.), *Time and behaviour: Psychological and neurobehavioural analyses* (pp. 79-131). Amsterdam: Elsevier.

Kristofferson, A. B. (1984). Quantal and deterministic timing in human duration discrimination. In J. Gibbon & L. Allan (Eds.), *Timing and time perception* (Annals of the New York Academy of Sciences, Vol. 423, pp. 3-15). New York: New York Academy of Sciences.

Luce, R. D. (1963). A threshold theory for simple detection experiments. *Psychological Review*, **70**, 61-79.

Marascuilo, L. A., & McSweeny, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Monterey, CA: Brooks.

Mattes, S., & Ulrich, R. (1998). Directed attention prolongs the perceived duration of a brief stimulus. *Perception & Psychophysics*, **60**, 1305-1317.

Miller, G. A., & Garner, W. R. (1944). Effect of random presentation on the psychometric function: Implications for a quantal theory of discrimination. *American Journal of Psychology*, **57**, 451-467.

Miller, J. [O.] (1998). Cupid: A program for computations with probability distributions. *Behavior Research Methods, Instruments, & Computers*, **30**, 544-545.

Miller, J. O., Patterson, T., & Ulrich, R. (1998). Jackknife-based method for measuring LRP onset latency differences. *Psychophysiology*, **35**, 99-115.

Mortensen, U. (1988). Visual contrast detection by a single channel versus probability summation among channels. *Biological Cybernetics*, **59**, 137-147.

Mortensen, U., & Suhl, U. (1991). An evaluation of sensory noise in the human visual system. *Biological Cybernetics*, **66**, 37-47.

Mühlbradt, L., Mattes, S., Möhlmann, H., & Ulrich, R. (1994). Touch sensitivity of natural teeth and endosseous implants revealed by difference thresholds. *International Journal of Oral & Maxillofacial Implants*, **9**, 412-416.

Nachmias, J. (1981). On the psychometric function for contrast detection. *Vision Research*, **21**, 215-223.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1986). *Numerical recipes: The art of scientific computing*. Cambridge: Cambridge University Press.

Purcell, D. G., & Stewart, A. L. (1988). The face-detection effect: Configuration enhances detection. *Perception & Psychophysics*, **43**, 355-366.

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

Rinkenauer, G., Mattes, S., & Ulrich, R. (1999). The surface–weight illusion: On the contribution of grip force to perceived heaviness. *Perception & Psychophysics*, **61**, 23-30.

Rosenbrock, H. H. (1960). An automatic method for finding the greatest or least value of a function. *Computer Journal*, **3**, 175-184.

Simpson, T. L. (1995). A comparison of six methods to estimate thresholds from psychometric functions. *Behavior Research Methods, Instruments, & Computers*, **27**, 459-469.

Spearman, C. (1908). The method of "right and wrong cases" ("constant stimuli") without Gauss's formulae. *British Journal of Psychology*, **2**, 227-242.

Spiegel, M. R. (1975). *Schaum's outline of theory and problems of probability and statistics*. New York: McGraw-Hill.

Sternberg, S., & Knoll, R. L. (1973). The perception of temporal order: Fundamental issues and a general model. In S. Kornblum (Ed.), *Attention and performance IV* (pp. 629-685). New York: Academic Press.

Sternberg, S., Knoll, R. L., & Zukofsky, P. (1982). Timing by skilled musicians. In D. Deutsch (Ed.), *The psychology of music* (pp. 181-239). New York: Academic Press.

Stevens, S. S., Morgan, C. T., & Volkmann, J. (1941). Theory of the neural quantum in the discrimination of loudness and pitch. *American Journal of Psychology*, **54**, 315-335.

Stuart, A., & Ord, J. K. (1987). *Kendall's advanced theory of statistics: Vol. 1. Distributional theory* (5th ed.). London: Charles Griffin.

Trevan, J. W. (1927). The error of determination of toxicity. *Proceedings of the Royal Society of London: Series B*, **101**, 483-514.

Ulrich, R. (1987). Threshold models of temporal-order judgments evaluated by a ternary response task. *Perception & Psychophysics*, **42**, 224-239.

Urban, F. M. (1907). On the method of just perceptible differences. *Psychological Review*, **14**, 244-253.

Woodworth, R. S., & Schlosberg, H. (1954). *Experimental psychology*. New York: Holt.

NOTES

1. In sensory psychophysics, a two-alternative forced-choice (2AFC) task is often used to measure absolute thresholds, instead of the classical yes–no task. In this task, the ordinate of the psychometric function ranges from 50% (chance performance) to 100% (perfect detection), and the absolute threshold is usually defined as the stimulus point at which 75% of all responses are correct. Although the function obtained in a 2AFC task is not strictly a psychometric function, because the percentage of responses does not range from 0% to 100%, methods for the analysis of psychometric functions can be adapted to this task, as we will describe in the General Discussion section.

2. The SK method (described in detail later) was used to obtain an estimate of skewness for each of the 30 observed psychometric functions reported in this study. A t test was then used to test the null hypothesis that the average skewness was zero.

3. Church and Cobb (1973) showed that the estimated mean $\hat{\mu}$ is a nonparametric maximum likelihood estimate for so-called *equal weight designs* when the Ayer et al. (1995) procedure is applied. For example, a typical psychophysical experiment with equally spaced stimulus levels and equal sample sizes at each stimulus level would be a special case of this design. More specifically, the SK estimate

$$\hat{\mu} = \frac{1}{2}\sum_{i=1}^{k+1}\left(s_i - s_{i-1}\right)\left(\overline{p}_i - \overline{p}_{i-1}\right) \tag{17}$$

of μ is a maximum likelihood estimate of

$$Q = \frac{1}{2}\sum_{i=1}^{k+1}\left(s_{i+1} - s_i\right)\left[F\left(s_i\right) - F\left(s_{i-1}\right)\right]. \tag{18}$$

This Q is the true mean of a distribution whose CDF is an approximation to that of the true psychometric function F. Specifically, this approximation CDF is equal to the true one at each stimulus value and is linearly interpolated between stimulus values.

4. Thus, there were actually only $27 \times 11 = 297$ trials per experiment in these simulations. In all simulations with 11 and 21 stimulus levels, the actual number of trials in the experiment was the nearest possible value allowing equal whole numbers of trials at each stimulus level. For example, there were $9 \times 11 = 99$ and $91 \times 11 = 1{,}001$ trials in the simulations with 11 stimulus levels and approximately 100 and 1,000 trials, respectively. The simulations with 5 stimulus levels included exactly the indicated numbers of trials, because these numbers were all even multiples of five. We thought that the constraint of equal numbers of trials per condition was more important than the small confounding of the number of stimulus levels with the total number of trials.

5. Further simulations with other placements of the most extreme stimulus values will be summarized in the General Discussion section.

6. A technical problem with the Naka–Rushton distribution is that only the first raw moment exists; higher raw moments do not. To circumvent this difficulty, we used a version of this distribution that was truncated at the value of 20, which has a CDF value of .9975. All raw moments exist for this truncated version of the distribution.

7. To introduce skew into the triangular distribution, the mode of the distribution was shifted away from the midpoint of the distribution. The Cupid program (J. O. Miller, 1998) was used to determine how far the mode needed to shift to produce each desired skewness value. The same program was also used to adjust the parameters of the mixture distributions, discussed next.

8. With 11 stimulus levels and 300 trials, for example, the proportions of samples excluded for each of the first six distributions listed in Table 3 were 2%, 18%, 91%, 7%, 74%, and 9%, respectively. Although the χ^2 test is generally recommended for use only with at least $n = 5$ samples per stimulus level, we ignored this recommendation in the simulations in which it was not met. We acknowledge that the χ^2 distribution provides a poorer approximation for determining significance in those simulations, but this detail seems to have had no major influence on the results.

9. We employed several checks to evaluate the accuracy of the numerical search algorithm. (1) The convergence criterion of the computer routine was set equal to 10^{-8}. Numerical examples confirmed that criterion values smaller than this value would only prolong the search, but clearly not affect the results obtained. (2) The search algorithm requires initial starting values for the parameters, and these were set to 1.0 and 0.4 for μ and σ, respectively. Additional simulations with different starting values—including the mean and standard deviation estimated by the SK method—and with restarting the routine at the claimed minimum (see Press, Flannery, Teukolsky, & Vetterling, 1986, p. 292) confirmed that the results were virtually unaffected by the choice of the initial starting value. (3) After generating 30,000 random samples from a given distribution and estimating parameters from each sample, we checked the resulting distribution of parameter estimates for outliers. Usually, none was found, indicating that the parameter search process did not occasionally wander into an extreme region of the parameter space. There was, however, one situation in which the numerical search algorithm yielded extremely unusual parameter estimates (e.g., estimated mean far outside the range of stimulus values): when the observed psychometric function tended to be decreasing rather than increasing. Because of binomial variability, this occasionally occurred in some of the simulations with smaller numbers of trials. Such deviant samples and their corresponding parameter estimates were excluded from all tabulations. It should also be mentioned that, in contrast, the SK method avoids these potential numerical problems, because parameter estimates are computed directly, instead of being obtained by a search process.

10. In simulated samples with relatively small numbers of trials per stimulus, it occasionally happened that the same response was given for all the trials at every stimulus level (i.e., $\hat{p}_i = 0$ or $\hat{p}_i = 1$ for all i). For a sample in which that happened, the bootstrap standard error is zero, because all the bootstrap samples were necessarily identical to each other (and to the observed sample). Such samples were excluded from all tabulations of results involving bootstrap standard errors. Note that the problem of zero estimated standard error would arise with any variability estimator using the observed sample proportions as the population estimates, because binomial variability equals zero if all population probabilities are either 0 or 1.

11. Because of the additional computational load imposed by bootstrapping, we carried out bootstrapping for only 3,000 data sets in each set of simulations—not for the full 30,000 from which we obtained parameter estimates. All of the reported summary statistics concerning bootstrap standard errors and confidence intervals are thus averaged over this smaller number of simulated data sets.

12. In this article, we report only selected summaries of the simulation results. Interested readers may obtain considerably more detailed tabulations of the simulation results in a computer-readable format from either author.

13. In the case of the SK-based estimator of the mean—and in fact, higher moments—it is possible to compute the bias of the estimator directly, rather than computing it via an average of simulation results. Specifically, $\mathrm{E}[\hat{\mu}'_r]$ can be computed directly from Equation 2 by using the expected value operator in conjunction with the fact that $\mathrm{E}[\hat{p}_i] = p_i$ for every i. Nonetheless, we report biases computed from the simulation results for the moment-based SK estimators to enhance comparisons of their biases against those of the other estimators, which must be computed from the simulations.

14. It would be possible for an estimator to have the lowest average bias only because it was much less biased with one distribution, even if it were slightly more biased with most or even all of the other distributions, and this would substantially weaken the case that this estimator was best. This situation did not arise in the present instance, however. We also examined nonparametric summaries across distributions based on ranking the different estimators for each distribution and then averaging the ranks, and the results of those summaries also favored the mean computed with the SK method. Similar analyses were also conducted for all other summary tables in this article (e.g., Table 9). The results of such nonparametric analyses always agreed with those obtained by simple averaging; therefore, only the averages are reported. Throughout this article, all averaging across distributions was done by weighting each distri-

bution equally, despite the fact that the distributions differed with respect to the proportions of samples passing the χ^2 test.

15. Observed values of location estimators can then be compared statistically either by combining the estimates across observers (e.g., via a *t* test) or by using estimates of the sampling variability of each observed estimate (e.g., estimated from bootstrapping). For the purposes of this section, we will ignore this step of the problem and merely seek to identify the most sensitive estimator on the basis of its single-sample properties.

16. Although it is unusual to compute d' values corresponding to estimators derived from psychometric functions, the underlying logic of the d' is appropriate with any random variable. In concept, the d' simply measures the separation between the distributions of two random variables, regardless of what is being measured, and thus indexes the power of a statistical test to discriminate between those two distributions (e.g., J. O. Miller, Patterson, & Ulrich, 1998). In this article, we use d' to estimate the power of both location and dispersion estimators to detect differences between two conditions.

17. A technical issue that arises in comparing estimators of the dispersion parameters σ and *dl* is that these estimators have slightly different scales. With the normal distribution, for example, $\sigma = 1.48 \times dl$. Therefore, estimators of σ would be expected to have larger standard errors than estimators of *dl*—and probably larger biases as well—simply because the values being estimated are larger. To correct for this, we computed the biases and standard errors of estimates of these parameters, relative to the actual values of the parameters. For example, the bias in the probit estimate of the standard deviation was computed as

$$\frac{E\left[\hat{\sigma}_{\mathrm{PR}}\right]-\sigma}{\sigma}.$$

After such normalization, estimates appear to be on comparable scales for both parameters.

18. We also examined the performance of bootstrap confidence intervals and the power of experiments to detect differences in skewness between two experimental conditions, but this information is omitted in the interests of brevity. Readers interested in this information should contact the first author for additional information. The same is true for the kurtosis estimators considered in the following section.

19. More generally, for an *m*-AFC task, the transformation is

$$\hat{p}_i = \frac{m \cdot \hat{g}_i - 1}{m - 1}.$$

20. This possibility was suggested by Stanley Klein.

APPENDIX
Equations for the Psychometric Functions Used in the Simulations

normal $F(s) = \Phi\left(\frac{s-0.5}{0.25}\right)$, where $\Phi(z)$ is the cumulative standard normal

quantal $F(s) = 1 - \sum_{i=0}^{c-1} \frac{s^i e^{-s}}{i!}$, where $c = 4$

Naka–Rushton $F(s) = \frac{s^2}{1+s^2}$

Weibull $F(s) = 1 - e^{-s^2}$

model 4 $F(s) = \begin{cases} \frac{1}{2 \cdot t} & \text{for } -t \le s \le 0 \text{ or } t \le s \le 2t \\ 0 & \text{elsewhere} \end{cases}$

triangular(–0.8) $F(s) = \begin{cases} 1.19 \cdot s^2 & 0 \le s \le 0.84 \\ 1 - 6.37 \cdot (1-s)^2 & 0.84 < s \le 1 \end{cases}$

triangular(–0.4) $F(s) = \begin{cases} 1.88 \cdot s^2 & 0 \le s \le 0.53 \\ 1 - 2.14 \cdot (1-s)^2 & 0.53 < s \le 1 \end{cases}$

triangular(0.0) $F(s) = \begin{cases} 2.00 \cdot s^2 & 0 \le s \le 0.50 \\ 1 - 2.00 \cdot (1-s)^2 & 0.50 < s \le 1 \end{cases}$

triangular(0.4) $F(s) = \begin{cases} 2.14 \cdot s^2 & 0 \le s \le 0.47 \\ 1 - 1.88 \cdot (1-s)^2 & 0.47 < s \le 1 \end{cases}$

triangular(0.8) $F(s) = \begin{cases} 6.37 \cdot s^2 & 0 \le s \le 0.16 \\ 1 - 1.19 \cdot (1-s)^2 & 0.16 < s \le 1 \end{cases}$

APPENDIX (Continued)

mixture(−0.8)	$F(s)=\begin{cases}0.500\cdot s & 0\le s\le 0.423\\ 0.212+1.367\cdot(s-0.423) & 0.423<s\le 1\end{cases}$
mixture(−0.4)	$F(s)=\begin{cases}0.500\cdot s & 0\le s\le 0.148\\ 0.074+1.087\cdot(s-0.148) & 0.148<s\le 1\end{cases}$
mixture(0)	$F(s)=1$ for $0\le s\le 1$
mixture(0.4)	$F(s)=\begin{cases}1.087\cdot s & 0\le s\le 0.852\\ 0.926+0.500\cdot(s-0.852) & 0.852<s\le 1\end{cases}$
mixture(0.8)	$F(s)=\begin{cases}0.500\cdot s & 0\le s\le 0.577\\ 0.289+2.318\cdot(s-0.577) & 0.577<s\le 1\end{cases}$
t(k)	Student's *t* distribution with *k* degrees of freedom.

(Manuscript received November 29, 2000;
revision accepted for publication August 15, 2001.)