Perception & Psychophysics
2004, 66 (3), 517-533

Threshold estimation in two-alternative forced-choice (2AFC) tasks: The Spearman–Kärber method

ROLF ULRICH
Universität Tübingen, Tübingen, Germany

and

JEFF MILLER
University of Otago, Dunedin, New Zealand

The Spearman–Kärber method can be used to estimate the threshold value or difference limen in two-alternative forced-choice tasks. This method yields a simple estimator for the difference limen and its standard error, so that both can be calculated with a pocket calculator. In contrast to previous estimators, the present approach does not require any assumptions about the shape of the true underlying psychometric function. The performance of this new nonparametric estimator is compared with the standard technique of probit analysis. The Spearman–Kärber method appears to be a valuable addition to the toolbox of psychophysical methods, because it is most accurate for estimating the mean (i.e., absolute and difference thresholds) and dispersion of the psychometric function, although it is not optimal for estimating percentile-based parameters of this function.

The estimation of sensory thresholds is a major issue in psychophysics. Despite the fact that psychophysical methods have been in use for more than 100 years (e.g., Fechner, 1860), the development of appropriate and efficient statistical methods is still in progress (see Klein & Macmillan, 2001). Indeed, a recent special issue of *Perception & Psychophysics*, edited by Klein and Macmillan, covered new developments in such methods.

A particularly useful method for measuring sensory thresholds is the so-called two-alternative forced-choice (2AFC) procedure, which was suggested by Blackwell (1953) for vision research and by Jones (1956) for research on taste and smell (see Engen, 1971; Gescheider, 1997). Psychophysicists often prefer this procedure over the classical yes–no task for determining thresholds, because the 2AFC procedure discourages response biases and also produces an especially high level of performance (e.g., Gescheider, 1997; Linschoten, Harvey, Eller, & Jafek, 2001; for a detailed review, see Macmillan & Creelman, 1991, chap. 5).

This research was supported by the Deutsche Forschungsgemeinschaft (Ul 116/6-2). We thank Lew Harvey and Stanley Klein for helpful comments on a previous version of this article. Correspondence concerning this article should be addressed to R. Ulrich, Abteilung für Allgemeine Psychologie und Methodenlehre, Psychologisches Institut, Universität Tübingen, Friedrichstr. 21, 72072 Tübingen, Germany, or to J. Miller, Department of Psychology, University of Otago, Dunedin, New Zealand (e-mail: ulrich@uni-tuebingen.de or miller@psy.otago.ac.nz).

Note—This article was accepted by the previous editorial team, headed by Neil Macmillan.

The 2AFC procedure can be used to estimate both absolute thresholds in detection tasks and difference thresholds in discrimination tasks. In each trial of a 2AFC detection task, for example, the participant observes two well-defined time intervals. One interval—randomly either the first or the second—contains a signal (e.g., a weak tone), and the participant knows that exactly one signal will be presented during each trial. At the end of the trial, the participant indicates whether the first or the second interval contained the signal, and the experimenter simply notes whether the response is correct. Thus performance is measured as the proportion of correct responses. This proportion varies from the chance level of .5 for very weak signals up to 1.0 for very strong ones. The detection threshold is usually defined as the stimulus intensity at which the proportion of correct responses is .75 (e.g., McKee, Klein, & Teller, 1985).

In a 2AFC discrimination task, two stimuli—a standard (S) and a comparison (C)—are presented one at a time in the two successive intervals, with the order of presentation varying randomly from trial to trial. The S and the C differ along a certain physical dimension, such as object weight, light intensity, or molar concentration (e.g., Luce, 1993). The C is always the more extreme stimulus on this dimension (i.e., heavier, higher intensity, etc.), and the value of the C along this dimension varies from trial to trial. The participant is asked to indicate which interval contained the more extreme stimulus, and performance is again summarized in terms of the proportion of correct responses, varying from .5 to 1.0. Note that the 2AFC design should not be confused with the so-called *reminder*

design (see Macmillan & Creelman, 1991), which is a version of the method of constant stimuli in which the S and the C are always presented in the same order. In the reminder design, the C may be *more extreme* or *less extreme* than the S. At the end of each trial, the participant judges whether the C was more or less extreme than S. Critically, in this case the psychometric function ranges from 0 to 1.0 rather than from .5 to 1.0, as in a 2AFC task.

For example, in a 2AFC discrimination task, S might be a standard weight of 100 g, and C might be a heavier comparison weight selected randomly in each trial from the following set of weights: 100, 102, 104, 106, 108, or 110 g. At the end of the trial, the participant indicates whether the first or the second stimulus was heavier. As in the detection task, the proportion of correct responses will vary from the chance level of .5 (when the two stimuli do not differ) to 1.0 (when the two stimuli are readily distinguishable). Analogous to detection tasks, the discrimination threshold is commonly defined as the stimulus magnitude of the comparison at which the proportion of correct responses is equal to .75.

There are also several variants of the 2AFC tasks (for an overview, see Macmillan & Creelman, 1991, chap. 5). For example, in a discrimination task, S and C might be presented simultaneously instead of successively. In a visual discrimination task, for example, they might be presented side by side, with the observer being asked to indicate which side contains the more extreme stimulus.

Figure 1 provides a hypothetical outcome for the 2AFC weight discrimination task described above. The abscissa of this figure represents the difference between the C and the S stimuli. For the above example, this difference would range from 0 to 10 g. The ordinate shows the proportion of correct responses, which, as mentioned before, increases from .5 to 1.0 as the weight of the C stimulus increases. The resulting psychometric function might be S-shaped (as is shown in the figure) or might have any other shape.[1] Apart from random fluctuation, however, these psychometric functions increase monotonically with increasing stimulus magnitude. Indeed, the new method proposed in this article for the analysis of 2AFC results requires only that the true underlying 2AFC psychometric function is monotonically increasing.

Several approaches have been suggested to summarize the participant's performance, given 2AFC data like those shown in Figure 1 (e.g., Klein, 2001; Leek, 2001). These approaches usually proceed explicitly from the assumption that a true but unobservable psychometric function $G(x)$ underlies performance in the 2AFC task. A general way of writing this true psychometric function is

$$G(x) = 0.5 + 0.5 \cdot F(x), \qquad (1)$$

where $F(x)$ represents a cumulative distribution function (CDF; e.g., Green, Richards, & Forrest, 1989; Harvey, 1986; Klein, 2001; Miller & Ulrich, 2001; Mortensen, 2002).[2] Usually, the researcher assumes a specific distributional form for $F(x)$. Most commonly, $F(x)$ is assumed to be the CDF of the normal, Weibull, logistic, or hyperbolic tangent distribution (e.g., Strasburger, 2001a). The function $G(x)$ merely rescales this CDF by introducing a correction for guessing, as is appropriate for the 2AFC task.

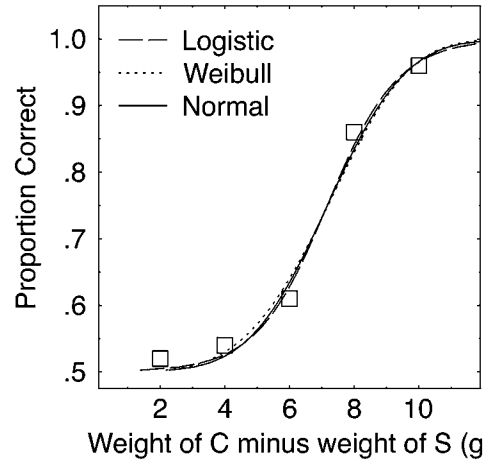

Figure 1. A typical psychometric function in a two-alternative forced-choice weight discrimination paradigm. It shows the probability of a correct response as a function of the difference x between the weights of the standard (S) and comparison (C) stimuli—in this case, for a standard stimulus of 100 g. The data were generated by a Monte Carlo simulation with a skewed triangular function having a range from 1 to 11 and a mode at 9 as the true underlying function $F(x)$. The mean and the standard deviation of this function were 7.00 and 2.16, respectively. Each of the six comparison stimuli was presented 200 times. The smooth lines show the best-fitting logistic, normal, and Weibull functions, which were estimated by a maximum-likelihood procedure as described in the text.

On the basis of this conceptualization of the 2AFC task, the researcher estimates the parameters of the distribution assumed for $F(x)$ and summarizes performance in terms of the estimated distribution. Traditionally, the threshold has been taken as the median of this distribution [i.e., the value of x for which $F(x) = .5$; see, e.g., Linschoten et al., 2001; Wichmann & Hill, 2001; for a review, see Gescheider, 1997]. As will be discussed further below, however, it is possible that the mean should be used instead of the median. The mean generally has better statistical properties than the median (e.g., Sen, 1985; Stuart & Ord, 1987), and the mean is known to be a more stable estimator than the median for psychometric functions in yes–no tasks (e.g., Church & Cobb, 1973; Miller & Ulrich, 2001). In addition, the mean may be more appropriate than the median for testing theoretical models that make predictions about the moments, rather than the percentiles, of psychometric functions (e.g., Mortensen, 2002; Sternberg & Knoll, 1973; Sternberg, Knoll, & Zukofsky, 1982; Ulrich, 1987). Of course, it does not matter whether the median or the mean is used if the assumed distribution $F(x)$ is symmetric, because in that case, these two values are identical.

For example, if one assumes a symmetric logistic CDF for $F(x)$, Equation 1 can be written as

$$G(x) = 0.5 + 0.5 \cdot \left[1 + e^{-\frac{x-\alpha}{\beta}}\right]^{-1}, \quad (2)$$

where α is the location parameter and $\beta > 0$ is the scale parameter. Note that the location parameter α is both the mean and the median of the logistic distribution, so this parameter is usually referred to as the *threshold* in studies assuming the logistic function (see Linschoten et al., 2001). The parameters α and β in Equation 2 can be estimated by the maximum likelihood method (e.g., Harvey, 1986; Treutwein & Strasburger, 1999). For the data shown in Figure 1, for example, the maximum-likelihood estimates are $\hat{\alpha} = 7.17$ and $\hat{\beta} = 1.09$.[3] The dashed line in this figure, then, represents the fitted function.

As a different example, if one assumes the asymmetric Weibull CDF for $F(x)$, Equation 1 can be written as

$$G(x) = 0.5 + 0.5 \cdot \left[1 - e^{-(\alpha \cdot x)^{\beta}}\right], \quad (3)$$

where $\alpha > 0$ and $\beta > 0$ are the scale and shape parameters, respectively, and $x \geq 0$ (e.g., Ross, 2000). For the data shown in Figure 1, for example, the maximum-likelihood estimates are $\hat{\alpha} = 7.86$ and $\hat{\beta} = 4.12$. On the basis of these parameter estimates, the median-based estimate of the threshold is 7.19, and the mean-based estimate is 7.13.

The preceding examples illustrate two important limitations of existing methods for analyzing psychometric function data obtained in 2AFC tasks. First, the researcher must specify a particular probability distribution to which the observed response probabilities will be fitted. In practice, it is often unclear which distribution to select, and the choice of distribution can have an important impact on the results. Second, if an asymmetric distribution is selected, the researcher must make an additional choice of whether to estimate the threshold with the median or the mean of that distribution. The median is chosen implicitly by most researchers, who simply adopt the standard definition of the threshold as the stimulus value yielding 75% correct performance in the 2AFC task. Some researchers, however, have questioned whether the standard definition of 75% provides the most reliable measure and, therefore, have suggested that other performance levels be used (e.g., Gescheider, 1997, pp. 164–165). Specifically, Green (1990) argued that the definition associated with the highest reliability is somewhere in the 84%–94% range. Furthermore, the standard 75% definition cannot easily be used with some adaptive procedures that target other performance levels, such as 70.7% (e.g., Leek, 2001) or 84% (e.g., García-Pérez, 2002).

The present article addresses both of these limitations. Our main goal was to show how the nonparametric Spearman–Kärber method (e.g., Epstein & Churchman, 1944; Kärber, 1931; Miller & Ulrich, 2001; Spearman, 1908) can be used for the analysis of 2AFC data, thereby eliminating the need for any assumption about the underlying distribution. The Spearman–Kärber method was developed for the analysis of data from the yes–no task, and Miller and Ulrich (2001) found that this method works quite well for the analysis of yes–no data. They showed that the method could, in principle, also be applied to data from forced-choice tasks, but they did not conduct any simulations to examine the accuracy of the method with forced-choice data. Moreover, Klein (2001, pp. 1439–1442) argued on several grounds that the Spearman–Kärber method was unlikely to perform nearly as well with forced-choice data as it did with yes–no data. For example, he reasoned that the Spearman–Kärber method gives equal weight to the observed probabilities at all stimulus levels, whereas parametric procedures, such as probit analysis, give lower weights to probabilities closer to .5, which have larger binomial variability. Hence, he argued that the Spearman–Kärber method would provide less reliable threshold estimates than would probit analysis. Therefore, one goal of the present article was to provide a more detailed assessment of the Spearman–Kärber method for the analysis of 2AFC data. Specifically, we examined the statistical properties of the Spearman–Kärber method and contrasted these with traditional estimation methods, such as probit analysis. Furthermore, as will be developed below, the Spearman–Kärber method provides different estimates of the median and the mean when the 2AFC data seem to reflect an asymmetric distribution. Thus, as a secondary goal, we sought to determine whether the median or the mean should be used, by comparing the statistical properties of these two estimators.

EXTENDING THE SPEARMAN–KÄRBER METHOD TO 2AFC TASKS

It is helpful to provide a brief review of the Spearman–Kärber method for the analysis of yes–no data before generalizing this method to 2AFC tasks. As was mentioned before, the ordinate $F(x)$ of the classical yes–no function ranges from 0 to 1 as stimulus magnitude x increases. This function is commonly assumed to be a nondecreasing function and, thus, can be regarded as a CDF from some probability distribution (Falmagne, 1985; Luce, 1963; Trevan, 1927). Probit analysis, for example, assumes that this CDF can be approximated by a normal ogive and estimates the mean and the standard deviation of the assumed underlying normal CDF from the observed psychometric function.

Like probit analysis and other estimation procedures, the Spearman–Kärber method treats the psychometric function observed in a yes–no task as a CDF[4] from which any desired percentiles or moments can be estimated. Specifically, suppose a researcher uses k monotonically increasing stimulus values, $x_1 < x_2 \cdots < x_k$, to determine the observed response probabilities, $\hat{p}_i$ ($i = 1, \ldots, k$), associated with each stimulus value.[5] Then it can be shown (Church & Cobb, 1973; Sternberg et al., 1982,

pp. 234–236) that the mean of the underlying psychometric function is given by[6]

$$\hat{\mu} = \frac{1}{2}\sum_{i=1}^{k+1}\left(\hat{p}_i - \hat{p}_{i-1}\right)\left(x_i + x_{i-1}\right). \quad (4)$$

The values x_0 and x_{k+1} are chosen such that one can assume $p_0 = 0$ and $p_{k+1} = 1$.

In order to apply the Spearman–Kärber method to the 2AFC task, the psychometric function $G(x)$ defined by Equation 1 can be rearranged to recover the CDF $F(x)$ (cf. Miller & Ulrich, 2001)

$$F(x) = 2 \cdot G(x) - 1. \quad (5)$$

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

$$\hat{p}_i = 2 \cdot \hat{g}_i - 1, \quad (6)$$

and these transformed values are needed for the Spearman–Kärber method. Klein (2001) suggested an alternative approach that can be used when observers have a strong response bias in 2AFC tasks, but for simplicity we will consider in this article only situations in which response biases are small.

As an illustration, consider once more the above weight discrimination example. The observed response probabilities were $\hat{g}$ = (.52, .54, .61, .86, .96) at the stimulus values x = (2, 4, 6, 8, 10).[7] Equation 6 yields $\hat{p}$ = (.04, .08, .22, .71, .92), and application of the standard Spearman–Kärber method with these $\hat{p}$ values gives a mean of 7.04 and, thus, provides a nonparametric estimate of the threshold value. This estimate can be directly computed, if Equation 6 is inserted into Equation 4—that is,

$$\hat{\mu}_{2\mathrm{AFC}} = \sum_{i=1}^{k+1}\left(\hat{g}_i - \hat{g}_{i-1}\right)\left(x_i + x_{i-1}\right). \quad (7)$$

Note that x_0 and x_{k+1} are only required for the calculation of $\hat{\mu}_{2\mathrm{AFC}}$ but are not actually included in the experimental design. It is advisable to let x_0 be equal to the smallest admissible value of x, so that p_0 corresponds to the guessing probability .5. The value x_{k+1} should be large enough that $g_{k+1} = 1$ can be assumed, and in general, the difference $x_{k+1} - x_k$ need not match the differences between other stimulus levels. If $\hat{g}_k = 1$, the selection of x_{k+1} does not influence the estimate $\hat{\mu}_{2\mathrm{AFC}}$. For the above example, reasonable values of x_0 and x_{k+1} would be 0 and 12, respectively.

In addition, it can be shown that the standard error associated with the threshold estimate $\hat{\mu}_{2\mathrm{AFC}}$ is (see Appendix A)

$$SE\left(\hat{\mu}_{2\mathrm{AFC}}\right) = \sqrt{\sum_{i=1}^{k}\frac{\hat{g}_i \cdot \left(1 - \hat{g}_i\right)}{n_i - 1} \cdot \left(x_{i+1} - x_{i-1}\right)^2}, \quad (8)$$

where n_i is the number of observations at stimulus level n_i. Expressions 7 and 8 become especially handy if the spacing between two successive stimulus levels is constant—that is, $d = x_i - x_{i-1}$ for $i = 1, \ldots, k + 1$. In that case, these expressions simplify to

$$\hat{\mu}_{2\mathrm{AFC}} = x_0 + 2 \cdot d \cdot \left[k + \frac{5}{4} - \sum_{i=1}^{k+1}\hat{g}_i\right] \quad (9)$$

and

$$SE\left(\hat{\mu}_{2\mathrm{AFC}}\right) = 2 \cdot d \cdot \sqrt{\sum_{i=1}^{k}\frac{\hat{g}_i \cdot \left(1 - g_i\right)}{n_i - 1}}, \quad (10)$$

respectively. For the above numerical example with 200 trials per stimulus level, this formula yields a standard error of 0.27.

SIMULATION METHOD

The evaluation of the different approaches for summarizing performance in the 2AFC task was similar to the one used by Miller and Ulrich (2001). In brief, the nonparametric Spearman–Kärber estimation procedure was compared with traditional parametric estimation procedures—that is, maximum-likelihood probit analysis. To this end, we generated simulated data in experiments in which the method of constant stimuli was used. The different sets of data were factorially varied according to the (1) number of stimulus levels (k = 5, 10, or 15), (2) total number of trials per psychometric function (N = 30, 60, 120, 240, or 480), and (3) true underlying psychometric function $F(x)$ (normal, logistic, Weibull, or hyperbolic tangent). The parameters for each distribution were adjusted with CUPID (Miller, 1998) so that the resulting mean and standard deviation of each distribution were 5 and 2, respectively. The medians of the normal and logistic functions were necessarily identical to the means, because these two functions are symmetric. The medians of the asymmetrical Weibull and hyperbolic tangent functions, however, were 4.908 and 4.509, respectively. The data set for each of the 60 factorial combinations was based on 30,000 simulated experiments. As in our previous simulations (Miller & Ulrich, 2001), the psychometric functions were monotonized before computing the estimates.[8]

It is helpful to consider in more detail the procedure and results for a single simulated experiment before we present the results averaged across experiments. As an example, consider a simulated experiment that used 10 stimulus levels, included 120 trials in total (i.e., 12 trials per stimulus level), and used the true underlying normal distribution $F(x)$ with mean $\mu = 5$ and standard deviation $\sigma = 2$. The stimulus levels were equally spaced, and the smallest and the largest stimulus levels x_1 and x_{10} were set to the 1st and the 99th percentiles of the true distribution—that is, $F(x_1) = .01$ and $F(x_{10}) = .99$. The resulting stimulus values, the associated probabilities $p_i = F(x_i)$, and the probability of a correct response $g_i = .5 + .5 \cdot p_i$ are shown in the first, second, and third rows of Table 1, respectively.[9] The fourth and fifth rows in this

Table 1
Stimulus Values x_i, Probabilities p_i and g_i, and Example Simulated Results for Simulations With Normal Psychometric Function, 10 Stimulus Levels, and 120 Experimental Trials

	Stimulus Level									
	1	2	3	4	5	6	7	8	9	10
	All Simulations									
x_i	0.35	1.38	2.42	3.45	4.48	5.52	6.55	7.59	8.62	9.65
$p_i = F(x_i)$	.01	.04	.10	.22	.40	.60	.78	.90	.97	.99
$g_i = G(x_i)$	.51	.52	.55	.61	.70	.80	.89	.95	.98	1.0
	One Simulation									
N of correct responses	8	3	9	8	9	11	10	12	12	12
N of incorrect responses	4	9	3	4	3	1	2	0	0	0
Estimated probability $\hat{g}_i$	.67	.25	.75	.67	.75	.92	.83	1.0	1.0	1.0

table contain the data of a single simulated experiment, giving the number of correct responses and the corresponding estimated probabilities $\hat{g}_i$ of a correct response at each stimulus level.

An identical analysis was performed on the results of each simulated experiment. The main steps of this analysis will be demonstrated for the simulated data set provided in Table 1 (for further details, see Miller & Ulrich, 2001). First, the Spearman–Kärber method was used to estimate the mean μ_{2AFC} and median med_{2AFC} of $F(x)$. Furthermore, we also estimated the standard deviation σ_{2AFC} and the difference limen dl_{2AFC} of $F(x)$, because these parameters are sometimes of interest in psychophysical research (e.g., Lam, Dubno, & Mills, 1999).

Second, the maximum-likelihood estimators of the mean, median, standard deviation, and difference limen were computed four times via probit analysis, separately under the assumptions that the underlying distribution could be approximated by the normal, the Weibull, the logistic, or the hyperbolic tangent function.

The fit of the simulated data set to each assumed function was evaluated by a χ^2 test (see Wichmann & Hill, 2001). If the computed χ^2 was significant at $p < .05$, the data set was regarded as inappropriate for this assumed function and was excluded from the overall tabulation of results for that assumed distribution. We discarded such data sets to mimic the procedure that would be followed in practice. Because a researcher does not know the true underlying psychometric function, he or she would tend to ignore estimates based on what appeared from the χ^2 test to be an inappropriate underlying function. The complete outcome of this analysis for the simulated data in Table 1 is summarized in Table 2.

SIMULATION RESULTS

For each of the 60 factorial combinations of simulation parameters and for each method of analysis, we computed the average across simulated experiments of the mean, median, standard deviation, and difference limen estimated from the data of each experiment (i.e., $\hat{\mu}_{2AFC}$, $\hat{med}_{2AFC}$, $\hat{\sigma}_{2AFC}$, and $\hat{dl}_{2AFC}$). In addition, we computed the standard deviation of each estimate across experiments to assess its standard error. For the parametric estimation methods assuming an underlying normal, logistic, Weibull, or hyperbolic tangent distribution, we included only simulated data sets that passed the χ^2 goodness-of-fit test in the computations of these averages and standard deviations. For the nonparametric Spearman–Kärber method, we included in these computations all simulated experiments within each factorial combination without regard to any χ^2 test.[10]

Table 3 shows the mean percentage of simulated experiments that passed the χ^2 goodness-of-fit test as a func-

Table 2
Estimated Parameters ($\hat{\mu}_{2AFC}$, $\hat{med}_{2AFC}$, $\hat{\sigma}_{2AFC}$, *and* $\hat{dl}_{2AFC}$) From the Simulated Data of Table 1

	Estimated Parameters				
Method of Analysis	$\hat{\mu}_{2AFC}$	$\hat{med}_{2AFC}$	$\hat{\sigma}_{2AFC}$	$\hat{dl}_{2AFC}$	χ^2
Spearman–Kärber	4.31	4.48	2.12	1.76	–
Normal	4.34	4.34	2.00	1.35	8.69
Logistic	4.32	4.32	2.13	1.29	8.91
Weibull	4.41	4.30	1.86	1.30	8.61
Hyperbolic tangent	4.18	3.60	2.36	1.45	8.26

Note—Each line provides the estimates for a different method of analysis (Spearman–Kärber and probit analysis with the normal, logistic, Weibull, and hyperbolic tangent functions). The last column gives the obtained χ^2 value for the goodness-of-fit test conducted for each distributional assumption used for probit analysis. In each case, the degrees of freedom of the χ^2 test are $df = 8$. In no case was the test significant (i.e., $p > .3$).

Table 3
Average Percentage of Simulated Experiments That Passed the χ^2 Test as a Function of Simulation Condition and Method of Analysis

		Method of Analysis			
Simulation Factor	Level	N	L	W	H
True distribution	N	98.4	98.5	98.4	97.3
	L	98.1	98.3	98.1	97.6
	W	98.2	98.2	98.5	97.9
	H	92.7	94.7	95.1	99.0
Total number of trials	30	99.7	99.7	99.7	99.6
	60	99.0	99.0	99.1	99.0
	120	97.8	97.9	98.3	98.2
	240	95.8	96.4	96.7	97.3
	480	92.1	94.2	93.8	95.6
Number of stimulus levels	5	94.3	95.4	95.4	96.5
	10	97.5	98.0	98.1	98.4
	15	98.8	99.0	99.1	98.9

Note—N, normal; L, logistic; W, Weibull; H, hyperbolic tangent.

tion of each simulation factor, and the results indicate that the test has limited usefulness for discriminating between different underlying distributions.[11] Consider first the results as a function of the true distribution. Ideally, 95% of the simulated experiments should pass the test when the method of analysis matches the true distribution (e.g., normal true distribution and normal-based method of analysis). In fact, 98%–99% of the simulated experiments passed the goodness-of-fit test when the analysis was based on the correct true distribution, indicating that the actual Type I error rate of the test in practice is slightly below the theoretically expected 5%, presumably due to the fact that the χ^2 test is only approximate. Also ideally, a much smaller percentage of simulated experiments should pass the goodness-of-fit test when the method of analysis does not match the true distribution (e.g., normal true distribution but logistic-based analysis). In fact, however, simulated experiments are just about as likely to pass the goodness-of-fit test whether the assumed distribution does or does not match the true one, indicating that the test has very little power to discriminate the true distribution from an alternative one under these conditions. The hyperbolic tangent distribution provides a partial exception to this generalization: Data generated from it fairly often failed the goodness-of-fit test associated with the fit of a normal, logistic, or Weibull model, although data from the latter models rarely failed the goodness-of-fit test associated with fitting the hyperbolic tangent model.

Consider next the performance of the test when the total number of trials is varied. In general, many more data sets pass the test when the number of trials is smaller, partly because the power of the test increases with the number of trials. Finally, the test tends to accept slightly more data sets when the number of stimulus levels is larger. These results further strengthen the conclusion reached by Wichmann and Hill (2001) that the traditional χ^2 test should be replaced by alternative methods that do not rely on large-sample theory.

Mean and Median

As was mentioned in the Method section, for each simulated condition and method of analysis, we computed the means and standard deviations of the estimates of the mean and median across all simulated experiments. Biases were then measured as the absolute values of the differences between the average estimated values and the true values. Absolute values of the bias were used because signed biases in opposite directions could cancel each other out when averaged. Note that the standard deviation of each estimator computed across simulations provides an estimate of its standard error.

The results for the mean and the median estimators were similar, except that the median estimators tended to be more biased and less reliable than the mean estimators. For example, the average bias of the probit estimators was 0.158 for the median and 0.108 for the mean. Likewise, the average standard error of the probit estimators was 0.803 for the difference limen and 0.755 for the standard deviation. An identical pattern of results was obtained for the Spearman–Kärber method. The simulation results suggest, then, that in practice the mean should always be preferred over the median as an estimate of central tendency. An overall advantage for the mean over the median is not surprising, because the statistical properties of the mean are known to be better than those of the median in most situations (e.g., Sen, 1985; Stuart & Ord, 1987). Because it was the superior estimator, we report below detailed results only for the mean. These results are consistent with our previous simulation results for the yes–no task (Miller & Ulrich, 2001), which also showed that the mean was less biased and more reliable than the median.

Bias of mean. Table 4 shows the average absolute bias of each estimation method at each level of each factor varied in the simulations. Several results are striking. First and most important, the average absolute bias of the Spearman–Kärber method is consistently the least of any of these methods of analysis. In fact, biases of the other

Table 4
Average Absolute Biases of the Estimated Means as a Function of Simulation Condition and Method of Analysis

		Method of Analysis				
Simulation Factor	Level	N	L	W	H	SK
True distribution	N	0.055	0.049	0.084	0.136	**0.004**
	L	0.067	0.066	0.062	0.163	**0.002**
	W	0.087	0.086	0.050	0.112	**0.004**
	H	0.249	0.254	0.106	0.088	**0.043**
Total number of trials	30	0.164	0.169	0.117	0.282	**0.019**
	60	0.140	0.138	0.088	0.090	**0.013**
	120	0.103	0.102	0.067	0.042	**0.013**
	240	0.087	0.086	0.057	0.089	**0.011**
	480	0.078	0.074	0.050	0.121	**0.011**
Number of stimulus levels	5	0.111	0.113	0.039	0.144	**0.030**
	10	0.124	0.127	0.078	0.120	**0.006**
	15	0.109	0.102	0.110	0.111	**0.005**

Note—N, normal; L, logistic; W, Weibull; H, hyperbolic tangent; SK, Spearman–Kärber. The smallest value in each row is printed in boldface.

methods tend to be approximately eight times larger than biases of the Spearman–Kärber method. Even when the assumed distribution underlying a parametric analysis matches the true distribution used to generate the data (e.g., normal data analyzed assuming a normal underlying function), the means estimated with the Spearman–Kärber method are much less biased than the means estimated with the parametric method. Clearly, then, a researcher wanting a minimally biased estimate of the location of a psychometric function from a 2AFC task would be strongly advised to use the mean estimated by the Spearman–Kärber method.

Second, all methods become less biased as the number of trials increases. As a result, the advantage for the Spearman–Kärber method over the other methods decreases at larger numbers of trials as the overall biases of all methods converge toward zero. Third, the different methods of analysis exhibit differential sensitivity to the number of stimulus levels. At least over the ranges examined here, analyses assuming a normal or a logistic underlying function do not depend systematically on the number of levels. In contrast, bias increases with the number of stimulus levels when the Weibull is assumed, and it decreases with the number of stimulus levels when the hyperbolic tangent is assumed and when the Spearman–Kärber method is used. We have no explanation for these effects of the number of stimulus levels on probit analysis when the Weibull or the hyperbolic tangent function is assumed. The decrease in bias with increasing numbers of stimulus levels for the Spearman–Kärber method seems intuitively reasonable, however. This method approximates the underlying function with a polygon, and a greater number of points on the polygon would naturally give a better approximation.

Standard error of mean. Table 5 shows the average standard errors of the mean estimators for the different simulation conditions. The standard error of the Spearman–Kärber estimator is, on the average, approximately 6%

Table 5
Average Standard Errors of the Estimated Means as a Function of Simulation Condition and Method of Analysis

		Method of Analysis				
Simulation Factor	Level	N	L	W	H	SK
True distribution	N	0.793	0.793	**0.723**	0.755	0.849
	L	0.811	0.796	**0.711**	0.775	0.916
	W	0.778	0.778	**0.705**	0.731	0.784
	H	0.797	0.777	0.676	0.677	**0.663**
Total number of trials	30	1.408	1.393	**1.282**	1.307	1.418
	60	1.031	1.028	**0.900**	0.925	1.011
	120	0.705	0.695	**0.607**	0.646	0.715
	240	0.483	0.474	**0.423**	0.458	0.504
	480	0.348	0.341	**0.308**	0.338	0.368
Number of stimulus levels	5	0.801	0.784	**0.705**	0.762	0.874
	10	0.780	0.776	**0.700**	0.720	0.778
	15	0.803	0.798	**0.708**	0.723	0.758

Note—N, normal; L, logistic; W, Weibull; H, hyperbolic tangent; SK, Spearman–Kärber. The smallest value in each row is printed in boldface.

larger than those of the parametric methods, indicating that it is slightly less reliable than the other estimators.[12] As was expected, the standard errors of all methods decrease as the number of trials increases. In fact, the standard errors decrease with approximately the square root of the number of trials. For the parametric methods, standard errors depend little on the number of stimulus levels. For the Spearman–Kärber method, however, the standard error decreases with a larger number of levels. With 15 levels, in fact, standard errors were smaller on average with the Spearman–Kärber method than with the usual normal-based probit analysis, although the Weibull-based analysis had the smallest standard errors of all.

Standard Deviation and Difference Limen

As was done with the mean and median estimators, for each simulated condition and method of analysis, we computed the means and standard deviations of the estimates of the standard deviation and difference limen across all simulated experiments. These values were used to compute the biases and standard errors of these dispersion estimators just as they were for the central tendency estimators considered in the previous section. In general, the percentile-based estimators (i.e., $\hat{dl}_{2AFC}$) had lower biases and standard errors than did the moment-based estimators (i.e., $\hat{\sigma}_{2AFC}$). Specifically, the average bias of the probit estimators was 0.156 for $\hat{dl}_{2AFC}$ and 0.199 for $\hat{\sigma}_{2AFC}$, and the corresponding average standard errors were 0.660 and 1.015, respectively. A smaller bias for $\hat{dl}_{2AFC}$ than for $\hat{\sigma}_{2AFC}$ was also obtained for the Spearman–Kärber method (0.175 vs. 0.185). In contrast to the probit estimators, the Spearman–Kärber method's estimate of the difference limen was less reliable than its estimate of the standard deviation; the corresponding average standard errors were 0.661 and 0.559, respectively. These are somewhat inappropriate comparisons between the difference limen and the standard deviation, however, because the standard deviation is larger than the difference limen in the first place (see Miller & Ulrich, 2001); in a normal distribution, for example, $\sigma = 1.47 \cdot dl$. Taking into account these scale differences between the standard deviation and the difference limen, the moment-based estimators of the Spearman–Kärber method and of the hyperbolic tangent probit analysis yielded clearly superior results. For this reason and in order to save space, we will report only the results of the moment-based estimators.

Bias. Table 6 shows the average biases of the standard deviation estimates as a function of the simulation conditions. Estimates obtained with the parametric methods were, on average, 8% more biased than those obtained with the Spearman–Kärber method, although the estimates obtained with the hyperbolic tangent method were the least biased of all. Surprisingly, the hyperbolic tangent method produced the least bias regardless of the true underlying distribution—for example, it even outperformed the normal-based probit analysis when the true underlying distribution was normal. Thus, the simulation results suggest that researchers needing a minimally biased estimator of the dispersion of a psychometric function should consider using this method to estimate it. Interestingly, bias decreased as the number of trials increased for all methods except the hyperbolic tangent, so this recommendation is especially strong when there are relatively few trials. Bias also tended to decrease as the number of stimulus levels increased, at least for the hyperbolic tangent and the Spearman–Kärber methods, suggesting that a relatively large number of stimulus levels should be preferred when one of these methods is to be used for the estimation of dispersion.

Standard error. Table 7 shows the average standard errors of the standard deviation estimates as a function of the simulation conditions. The Spearman–Kärber method generally produces the smallest standard errors, although the Weibull method outperforms it slightly when the number of trials is large. The performance of the normal and the logistic methods is especially disappointing

Table 6
Average Absolute Biases of the Estimated Standard Deviations as a Function of Simulation Condition and Method of Analysis

		Method of Analysis				
Simulation Factor	Level	N	L	W	H	SK
True distribution	N	0.192	0.145	0.325	**0.094**	0.175
	L	0.238	0.187	0.391	**0.169**	0.193
	W	0.134	0.129	0.240	**0.053**	0.162
	H	0.255	0.259	0.228	**0.143**	0.210
Total number of trials	30	0.403	0.328	0.697	**0.123**	0.321
	60	0.217	0.174	0.344	**0.108**	0.187
	120	0.153	0.139	0.191	**0.097**	0.148
	240	0.132	0.129	0.137	**0.113**	0.138
	480	0.119	0.130	**0.112**	0.133	0.132
Number of stimulus levels	5	0.203	0.190	0.308	**0.182**	0.244
	10	0.204	0.175	0.281	**0.087**	0.127
	15	0.207	0.176	0.300	**0.076**	0.184

Note—N, normal; L, logistic; W, Weibull; H, hyperbolic tangent; SK, Spearman–Kärber. The smallest value in each row is printed in boldface.

Table 7
Average Standard Errors of the Estimated Standard Deviations as a Function of Simulation Condition and Method of Analysis

		Method of Analysis				
Simulation Factor	Level	N	L	W	H	SK
True distribution	N	0.975	1.083	0.905	1.159	**0.584**
	L	0.999	1.101	0.881	1.167	**0.635**
	W	0.964	1.075	0.872	1.063	**0.532**
	H	1.112	1.146	0.856	0.895	**0.485**
Total number of trials	30	1.962	2.142	1.729	2.020	**0.861**
	60	1.328	1.449	1.228	1.393	**0.687**
	120	0.848	0.923	0.718	0.915	**0.527**
	240	0.548	0.589	0.429	0.605	**0.405**
	480	0.378	0.402	**0.288**	0.422	0.314
Number of stimulus levels	5	0.995	1.098	0.870	1.097	**0.591**
	10	1.004	1.085	0.867	1.042	**0.546**
	15	1.039	1.121	0.899	1.074	**0.541**

Note—N, normal; L, logistic; W, Weibull; H, hyperbolic tangent; SK, Spearman–Kärber. The smallest value in each row is printed in boldface.

given that these are fairly standard methods. Their standard errors tend to be approximately double those of the Spearman–Kärber method, even when the true distribution matches the assumed one, until the number of trials gets rather large. As was expected, the standard errors of all of the estimators decrease as the number of trials increases. The number of stimulus levels has little effect on the standard errors.

EFFECTS OF STIMULUS LOCATIONS

In the simulations presented above, we assumed that the stimulus values covered almost the entire range of the 2AFC psychometric function from .5 to 1.0. More specifically, the smallest and largest levels were always set to the 1st and 99th percentiles of the true distribution—that is, $F(x_1) = .01$ and $F(x_k) = .99$, respectively. It is, however, possible that the Spearman–Kärber method would be much less effective if the stimulus series $x = (x_1, \ldots, x_k)$ was not broad enough to cover the whole transition zone of the true psychometric function (i.e., $p_1 >> 0$ or $p_k << 1$; see Woodworth & Schlosberg, 1954, p. 209). Although preliminary pilot testing of the stimuli may reduce the risk of truncation errors, it seems important to assess how such errors would influence threshold estimates.

For this purpose, we conducted further simulations with different combinations of truncation at the lower and upper tails of the true CDF. In particular, the percentiles $F(x_1) = (.01, .05, .10)$ at the lower tail were factorially combined with the percentiles $F(x_k) = (.99, .95, .90)$ at the upper tail. The normal distribution from the previous simulations served as the underlying CDF for this additional set of simulations. All the simulations were conducted with $k = 5$ stimulus levels and with 120 observations divided equally across levels. All the levels were equally spaced in terms of the z values of the underlying normal distribution, and the values of x_0 and x_6 were always placed one equal z-score step below x_1 and above x_5, respectively. Because the previous simulation results clearly suggest that the Spearman–Kärber estimate of the mean has to be favored over the Spearman–Kärber estimate of the median, we present the results for the Spearman–Kärber estimate of the mean only. In addition, we contrast these results with the results from traditional probit analysis.

Figure 2 summarizes the results from these simulations. The upper left panel indicates that, as was expected, the Spearman–Kärber estimator is most biased when the lower and upper stimuli are placed most asymmetrically [e.g., at (.01, .10) or at (.10, .01)]. Even in these worst cases, however, the absolute biases are still smaller than those of the probit estimator, as is shown in the upper right panel. Thus, these results extend the conclusion that the Spearman–Kärber mean is less biased than the probit mean to the case of moderately asymmetric stimulus placement. Interestingly, biases of the probit mean seem little affected by asymmetry of stimulus placement.

The standard errors of the mean estimators, shown in the lower two panels of the figure, reveal a strong tendency for means to be estimated more reliably (i.e., with smaller standard error) for more narrow stimulus placements. For the Spearman–Kärber method, truncation has a substantially larger effect for x_1 than for x_5, whereas the effects are more equal in magnitude for probit analysis. Perhaps most dramatic is the tendency for the standard error of the Spearman–Kärber method to improve more than that of probit analysis with increasing truncation. As was noted earlier, for example, the standard error is approximately 10% larger for the Spearman–Kärber estimates than for the probit estimates with the extreme placements of (.01, .99). In contrast, the Spearman–Kärber method has slightly smaller standard error than does the probit method with the extreme placements of (.05, .95), and it has much smaller standard error than does the probit method with the extreme placements of (.10, .90).

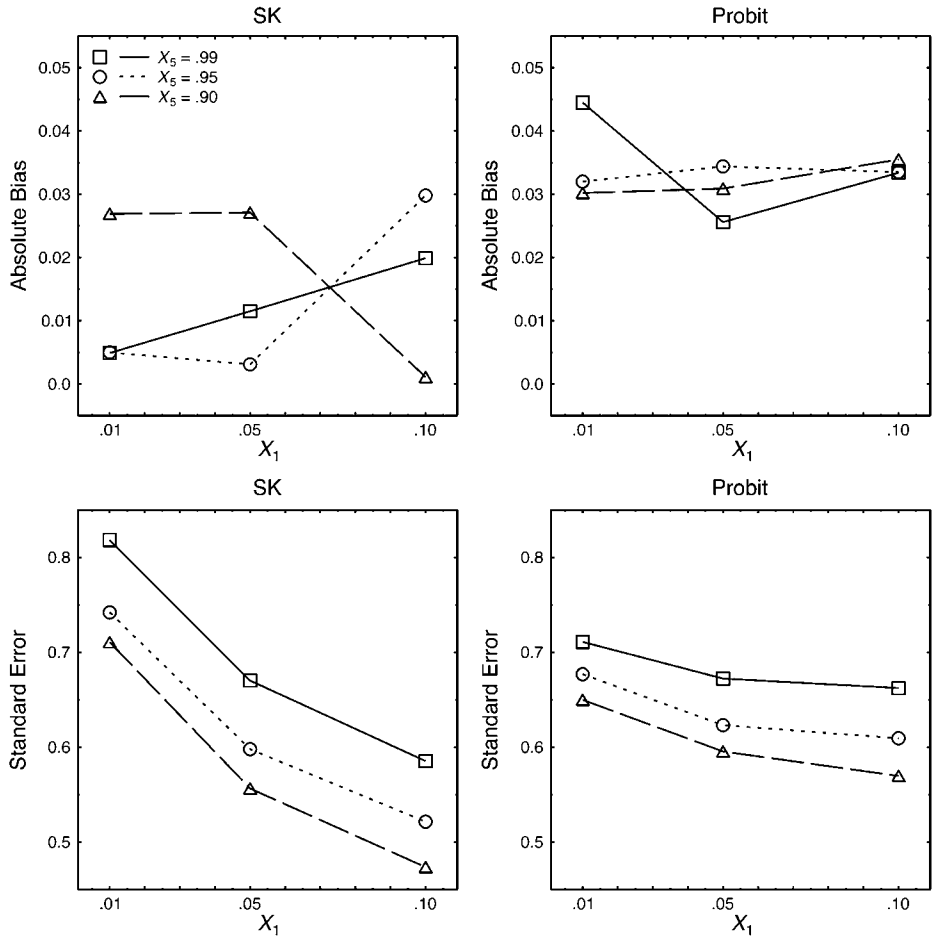

Figure 2. Effects of smallest (x_1) and largest (x_5) stimulus placements on the absolute bias of the Spearman–Kärber and probit mean estimators and their standard errors. The placement of each stimulus is specified in terms of the proportion truncated from the lower or upper tail of the true function $F(x)$. Upper left panel: bias of the Spearman–Kärber estimator. Upper right panel: bias of the probit estimator. Lower left panel: standard error of the Spearman–Kärber estimator. Lower right panel: standard error of the probit estimator.

Taking the bias results into consideration too, then, the Spearman–Kärber method is clearly quite a bit superior to the probit method with the stimulus placements examined here.

SUMMARY AND RECOMMENDATIONS

These simulations have examined the problem of estimating the location and dispersion of a psychometric function from 2AFC data, extending the work of Miller and Ulrich (2001) with the yes–no task. We compared five estimation procedures. Four used the probit method with different assumed underlying distributions (normal, logistic, Weibull, and hyperbolic tangent), and the fifth was the nonparametric Spearman–Kärber method. Across simulations, we varied the true underlying distribution from which the data were taken, the number of trials included in the experiment, and the number of stimulus levels at which the psychometric function was tested.

Table 8 summarizes the results. For each type of parameter, it shows the estimator that performed the best with respect to the two standard criteria—that is, minimizing bias and standard error. Two main results suggest that researchers should consider the Spearman–Kärber

Table 8
Most Successful 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*
Location	SK mean	Weibull mean
Dispersion	hyperbolic tangent *SD*	SK *SD*

Note—SK, Spearman–Kärber.

method as an alternative or addition to the estimation techniques that are regarded as standard in the current literature.

First, the Spearman–Kärber method appears to be a valuable addition to the toolbox of psychophysical methods for estimating parameters of a 2AFC psychophysical function. Its estimation performance keeps up with traditional methods, such as the probit analysis, although in contrast to classical methods, the Spearman–Kärber method does not require any assumption about the true underlying psychometric function. Second, the moment-based estimators consistently outperform the percentile-based Spearman–Kärber estimators. Although percentile-based estimators have a long tradition in psychophysical research, it appears that they should be largely discarded because they have clearly inferior statistical properties.

The rarely used Spearman–Kärber method clearly merits serious consideration as a new standard general-purpose estimation procedure. This method yields the least biased estimates of the mean and the lowest variance estimates of the standard deviation. It also yields reasonably unbiased estimates of the standard deviation and would be a good second choice for that purpose, after the hyperbolic tangent method. The most serious problem with the Spearman–Kärber method is that its estimates of the mean have approximately 9% higher standard errors than do those of the other estimators, although its reduced biases largely compensate for this increased variance with respect to the mean square error of estimation. In addition, the standard error can always be reduced to any desired level by increasing the size of the experiment. Finally, the Spearman–Kärber method does not require a χ^2 goodness-of-fit test before it can be applied to a given data set. Given the poor performance of the χ^2 test (Wichmann & Hill, 2001), this is an important advantage of the nonparametric procedure.

DISCUSSION AND POSSIBLE EXTENSIONS

The 2AFC task considered in this article and suggested by Blackwell (1953) and Jones (1956) is an extremely important and popular tool in psychophysical research (see Macmillan & Creelman, 1991). This is because its measured thresholds are not contaminated by certain response biases—that is, biases toward yes or no responses as in classical yes–no tasks—simply because *yes* and *no* are no longer possible response alternatives.[13] Furthermore, the 2AFC task yields relatively high levels of performance and, thus, appears to be an especially sensitive psychophysical tool (Macmillan & Creelman, 1991, p. 134).

It is therefore important to employ optimal statistical methods in the analysis of 2AFC data. The methods studied here are appropriate when the 2AFC task is used to generate a psychometric function to assess discrimination or detection performance. Although such psychometric functions are usually generated by the method of constant stimuli (e.g., Yeshurun, 1999), they can also be generated by adaptive procedures (e.g., Rinkenauer, Mattes, & Ulrich, 1999; Sternberg et al., 1982). It remains to be seen, however, how well the method will do when used with data obtained from adaptive psychophysical procedures (Klein, 2001).

The present simulation results suggest that the mean μ_{2AFC} should be preferred as an estimate of overall discrimination or detection sensitivity. Although the mean of a psychometric function is somewhat difficult to conceptualize, it actually has a fairly simple geometric interpretation in 2AFC tasks. As is shown in Appendix B, the mean equals half of the area bounded below by the psychometric function and above by a line corresponding to perfect performance at every stimulus level. Thus, this area must decrease as discrimination or detection sensitivity increases.

The present results suggest that the Spearman–Kärber method should be strongly considered as an estimator for the mean μ_{2AFC}. When the experimental requirements emphasize minimal bias, the Spearman–Kärber method is better than any of the probit-based estimators. Moreover, if the experimenter has succeeded in finding extreme stimulus values corresponding to proportions correct of approximately .525 and .975, the Spearman–Kärber method's estimator has the lowest standard error. When even more extreme stimulus values have been used (e.g., corresponding to proportions correct of approximately .505 and .995), however, the Spearman–Kärber method is probably not the best estimator when the experimental requirements emphasize minimal standard error.

The fact that the Spearman–Kärber method is often superior to probit analysis is perhaps somewhat surprising because, as was noted by Klein (2001), the Spearman–Kärber method weights all observed probabilities equally. In contrast, probit analysis gives less weighting to probability estimates with greater inherent binomial variability (Klein, 2001), as seems intuitively appropriate. Given the present simulation results, however, it appears that the advantages associated with the Spearman–Kärber method make it an effective estimation procedure for the 2AFC task with constant stimuli.

Interestingly, the Spearman–Kärber method also has implications for the design of experiments. Using Equation 8, for example, one could consider the question of how many trials should be tested at each stimulus level. In this article, we proceeded from the common practice that the researcher employs an equal number of trials at each stimulus level (i.e., $n_1 = n_2 = \cdots = n_k$). It is shown in Appendix C, however, that more reliable estimates of μ_{2AFC} can be obtained by using a somewhat greater number of trials at small than at large stimulus values (i.e., $n_1 > n_2 > \cdots > n_k$). In addition, this appendix provides a formula that could guide a researcher's choice in selecting an optimal adjustment of the number of trials n_i for each stimulus level x_i (see Bush, 1963). Interestingly, these recommendations based on a constant stimulus proce-

dure analyzed with the Spearman–Kärber method appear to conflict with recommendations based on adaptive procedures (Green, 1990; Klein, 2001), but as will be discussed in the Appendix, this conflict is actually illusory.

A related and important question concerns the issue of how many stimulus levels should be employed. The simulation results for the Spearman–Kärber method showed that the standard error of the mean decreased as the number of stimulus levels increased, even though the total number of trials remained constant. Therefore, the question arises whether this feature holds in general for the Spearman–Kärber method. The formal analysis provided in Appendix D shows that the estimate $\hat{\mu}_{2AFC}$ should indeed become more reliable in general as the number of stimulus levels is increased. Thus, when a researcher uses the Spearman–Kärber method, it seems advisable to employ many stimulus levels with a few trials per level, rather than a few stimulus levels with many trials per level. The gain from an increased number of stimulus levels is, however, subject to diminishing returns. For example, the gain in reliability will be greater when the number of levels is doubled from, say, 5 to 10 than from 10 to 20. In addition to the increased reliability of $\hat{\mu}_{2AFC}$, the simulations have clearly shown that bias of $\hat{\mu}_{2AFC}$ diminishes when a relatively large number of stimulus levels are employed. Thus, both reliability and bias considerations suggest that it is advisable to use many stimulus levels, if this is technically possible (e.g., 10 or more).

Another common design issue concerns the question of how stimulus levels should be spaced (e.g., Bush, 1963; Klein, 2001; Lam et al., 1999). Although it is common practice to space the stimulus levels equally, one might wonder whether an unequal spacing might produce more reliable estimates. For example, it might be more effective to lengthen the distance $d_i = x_i - x_{i-1}$ between two adjacent stimulus levels in a geometrically increasing fashion (e.g., $d_2 = a \cdot d_1, d_3 = a \cdot d_2, \ldots, d_k = a \cdot d_{k-1}$ with $a > 0$), rather than keeping d_i constant (e.g., $d_1 = d_2 = \cdots = d_k$). Intuitively, the geometrical spacing seems advantageous when the psychometric function $G(x)$ increases in a negatively accelerated fashion from the chance level to its asymptote. In such a case, the informative parts of the psychometric function would be traced more effectively with a geometrically increasing stimulus spacing than with a linear one. Again, the explicit expressions for $\hat{\mu}_{2AFC}$ and $\mathrm{Var}(\hat{\mu}_{2AFC})$ provided in this article might be helpful in deciding whether a linear or a geometrical spacing is appropriate.

In this article, we investigated the statistical properties of the nonparametric Spearman–Kärber method for the analysis of data from the 2AFC task. A previous study had shown that this method has superior properties for the analysis of data from yes–no tasks (Miller & Ulrich, 2001), but it was not clear whether the method would also work well with 2AFC tasks (e.g., Klein, 2001, pp. 1439–1442). Our results show that the Spearman–Kärber method does perform well when used with 2AFC data; that is, it provides estimates with less bias than do parametric estimators and with almost equal reliability. In sum, then, the present analysis shows that the Spearman–Kärber method might be preferred to parametric techniques for the analysis of psychometric functions obtained in 2AFC tasks, especially in situations in which an unbiased estimate of the threshold is desirable.

REFERENCES

Blackwell, H. R. (1953). *Psychophysical thresholds: Experimental studies of methods of measurement* (Rep. No. 36). Ann Arbor: University of Michigan, Bulletin of the Engineering Research Institute.

Bush, R. R. (1963). Estimation and evaluation. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. I, pp. 429-469). New York: Wiley.

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.

Engen, T. (1971). Psychophysics: I. Discrimination and detection. In J. W. Kling & L. A. Riggs (Eds.), *Woodworth and Schlosberg's experimental psychology* (3rd ed., pp. 11-46). New York: Holt, Rinehart & Winston.

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.* Leipzig: Breitkopf & Härtel.

Feller, W. (1971). *An introduction to probability theory and its applications* (Vol. II, 2nd ed.). New York: Wiley.

García-Pérez, M. A. (2002). Properties of some variants of adaptive staircases with fixed step sizes. *Spatial Vision,* **15**, 303-321.

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

Goldstein, L. J., Lay, D. L., & Schneider, D. I. (1987). *Calculus and its applications* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

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

Green, D. M., Richards, V. M., & Forrest, T. G. (1989). Stimulus step size and heterogeneous stimulus conditions in adaptive psychophysics. *Journal of the Acoustical Society of America,* **86**, 629-636.

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

Jones, F. N. (1956). A forced-choice method of limits. *American Journal of Psychology,* **69**, 672-673.

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.

Klein, S. A. (2001). Measuring, estimating, and understanding the psychometric function: A commentary. *Perception & Psychophysics,* **63**, 1421-1455.

Klein, S. A., & Macmillan, N. A. (Eds.) (2001). Psychometric functions and adaptive methods [Special issue]. *Perception & Psychophysics,* **63**(8).

Lam, C. F., Dubno, J. R., & Mills, J. H. (1999). Determination of optimal data placement for psychometric function estimation: A computer simulation. *Journal of the Acoustical Society of America,* **106**, 1969-1976.

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

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

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

Luce, R. D. (1993). *Sound and hearing: A conceptual introduction.* Hillsdale, NJ: Erlbaum.

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

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

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

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

Miller, J., & Ulrich, R. (2003). A computer program for Spearman–Kärber and probit analysis of psychometric function data. *Behavior Research Methods, Instruments, & Computers,* **36**, 11-16.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics* (3rd ed.). New York: McGraw-Hill.

Mortensen, U. (2002). Additive noise, Weibull functions, and the approximation of psychometric functions. *Vision Research,* **42**, 2371-2393.

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.

Ross, S. M. (2000). *Introduction to probability models* (7th ed.). San Diego: Academic Press.

Sen, A. R. (1985). Location parameter, estimation of. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 5, pp. 101-105). New York: Wiley.

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

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.

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

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

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

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

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.

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

Wonnacott, T. H., & Wonnacott, R. H. (1977). *Introductory statistics* (3rd ed.). New York: Wiley.

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

Yeshurun, Y. (1999). Spatial attention improves performance in spatial tasks. *Vision Research,* **39**, 293-306.

NOTES

1. Signal detection theory has been employed to analyze performance in 2AFC tasks (see Macmillan & Creelman, 1991, chap. 5). The signal detection model for 2AFC posits that both the S and the C generate a value along an internal continuum. These values are subject to random noise and thus a separate probability distribution is associated with each of the two stimuli. In each trial, the observer bases his response on the magnitude of these two values. Specifically, he will classify the first stimulus as the C, if the value generated by the first stimulus is larger than the value generated by the second stimulus, and vice versa. Under this assumption, the proportion of correct responses can be directly related to d'—that is, to the separation of the distributions associated with the two stimuli.

2. Some authors have included a lapsing rate, which describes imperfect performance (e.g., Harvey, 1986; Strasburger, 2001a). Because this rate is normally negligible, it is usually excluded from consideration for simplicity (e.g., Linschoten et al., 2001; Mortensen, 2002; Strasburger, 2001a, 2001b).

3. All of the maximum-likelihood estimates reported in this article were obtained with the PMETRIC computer program (Miller & Ulrich, 2003), which uses the simplex function minimization algorithm (Rosenbrock, 1960).

4. Observed psychometric functions are sometimes nonmonotonic. Nonmonotonicity arises because of binomial variability of the estimated response probabilities, especially in experiments with many stimulus values and only a few trials per stimulus value. When an observed psychometric function is nonmonotonic, it can be monotonized before the Spearman–Kärber method is applied. An algorithm to monotonize an observed function is described in Miller and Ulrich (2001), and a computer program for that purpose is provided by Klein (2001).

5. There is a slight ambiguity involving the scale of the x values. In detection tasks, these values are typically given in the appropriate absolute units of stimulus intensity (e.g., cd/m^2). In 2AFC discrimination tasks, however, the x values are sometimes specified in absolute units but sometimes specified as a difference in absolute units between the standard and the comparison. Fortunately, it is usually easy to tell from the context whether the x values are absolute or differences.

6. Measures of central tendency, such as the median or the mean, are the most important statistics for estimating the threshold value in a forced-choice situation. Therefore, we do not present explicit expressions for estimating higher moments, although we do present some simulation results concerning such estimates. Readers interested in the formulas for estimating higher moments should follow the computational steps given by Miller and Ulrich (2001, p. 1416).

7. Note that the observed response probabilities were rounded to two decimal places. This rounding affects slightly the following results of the example computations. For instance, without rounding error, the estimated mean $\hat{\mu}_{2AFC}$ is computed as 7.10 instead of 7.04.

8. It should be stressed, however, that monotonic psychometric functions are required only for the estimation of higher moments. The estimated mean is identical whether it is computed from the originally observed psychometric function or from the corresponding monotonized function as long as the stimulus levels are equidistant (see Sternberg et al., 1982, p. 235).

9. Because equidistant stimulus levels are common in psychophysical research, we did not evaluate the more complicated case with nonequidistant levels. In all the simulations, the values x_0 and x_{k+1} were set equal to $x_0 = x_1 - d$ and to $x_k + d$, respectively. Note that the constant d denotes the difference between two adjacent stimulus values in the experiment.

10. Due to binomial variability, a few simulated experiments yielded pathological psychometric functions. Such functions did not increase overall, and therefore, the maximum-likelihood procedure for estimating an increasing psychometric function did not converge to a finite estimate of the standard deviation. As one would expect, such pathological functions were observed more often when the number of total trials was small than when it was large. The proportions of pathological functions were .00425, .00008, .00000, .00000, and .00000 for 30, 60, 120, 240, and 480 total trials per psychometric function, respectively. These simulated experiments were excluded from all further analysis.

11. In this and subsequent analyses, we report the main effects of the simulation factors averaged across levels of the other factors, but we do not report interaction effects. This greatly simplifies the presentation of results but discards little important information, because the interaction effects tended to (1) be small relative to the main effects and (2) usually involved the effects of one factor's changing in size but not direction across levels of the other factors. Readers wishing to see the full analyses, including interactions as well as main effects, can obtain them from either author.

12. In the statistical literature, the relative efficiency of a biased estimator and an unbiased estimator is usually computed as the ratio of the

mean square errors of the two estimators. The mean square error of the unbiased estimator is simply its variance, whereas the mean square error of the biased estimator is its variance plus the square of its bias (Wonnacott & Wonnacott, 1977). In terms of statistical efficiency, then, the lower bias of the Spearman–Kärber method (see Table 4) compensates to some extent for its larger standard error, so that the efficiencies of the different estimators are approximately equal.

13. Although participants in a 2AFC task tend to exhibit less extreme biases, it should be stressed that this task cannot eliminate all biases. For example, in a temporal 2AFC task, in which the two stimuli are separated by time rather than by space, the classical time order error may still bias the results (temporal bias). Similarly, a participant may be biased toward one side when two stimuli are presented simultaneously side by side (positional bias).

APPENDIX A
The Standard Error of $\hat{\mu}_{2AFC}$

In order to derive the standard error of $\hat{\mu}_{2AFC}$, it will be helpful to rewrite the expression as

$$\hat{\mu}_{2AFC} = \sum_{i=1}^{k+1} \left(\hat{g}_i - \hat{g}_{i-1}\right) \cdot \left(x_i + x_{i-1}\right), \tag{A1}$$

which is Equation 7 from the main text. If we define $a_i = x_i + x_{i-1}$, the preceding expression can be rewritten as

$$\hat{\mu}_{2AFC} = \sum_{i=1}^{k+1} \hat{g}_i \cdot a_i - \sum_{i=1}^{k+1} \hat{g}_{i-1} \cdot a_i \tag{A2}$$

and further rearranged to

$$\hat{\mu}_{2AFC} = \hat{g}_{k+1} \cdot a_{k+1} - \hat{g}_0 \cdot a_1 + \sum_{i=1}^{k} \hat{g}_i \cdot \left(a_i - a_{i+1}\right). \tag{A3}$$

Since $\hat{g}_0 = .5$ and $\hat{g}_{k+1} = 1$ must hold by definition, this expression reduces to

$$\hat{\mu}_{2AFC} = a_{k+1} - \frac{a_1}{2} + \sum_{i=1}^{k} \hat{g}_i \cdot \left(a_i - a_{i+1}\right). \tag{A4}$$

Replacing $(a_i - a_{i+1})$ by $(x_{i-1} - x_{i+1})$ gives

$$\hat{\mu}_{2AFC} = a_{k+1} - \frac{a_1}{2} - \sum_{i=1}^{k} \hat{g}_i \cdot \left(x_{i+1} - x_{i-1}\right). \tag{A5}$$

Note that the variance of a linear combination $S = c_1 \cdot X_1 + \cdots + c_n \cdot X_n + c_{n+1}$ of independent random variables $X_1, \ldots, X_n$ with constants $c_1, \ldots, c_{n+1}$ is given by (e.g., Mood, Graybill, & Boes, 1974, p. 179)

$$\mathrm{Var}(S) = \sum_{i=1}^{n} c_i^2 \cdot \mathrm{Var}\left(X_i\right). \tag{A6}$$

Therefore, the variance of $\hat{\mu}_{2AFC}$ is obtained by Equation A5 as

$$\mathrm{Var}\left(\hat{\mu}_{2AFC}\right) = \mathrm{Var}\left[\sum_{i=1}^{k} \hat{g}_i \cdot \left(x_{i+1} - x_{i-1}\right)\right] \tag{A7}$$

$$= \sum_{i=1}^{k} \mathrm{Var}\left(\hat{g}_i\right) \cdot \left(x_{i+1} - x_{i-1}\right)^2, \tag{A8}$$

where $\mathrm{Var}(\hat{g}_i)$ is the variance of $\hat{g}_i$. Thus an unbiased estimator for this variance is provided by

$$\mathrm{Var}\left(\hat{g}_i\right) = \frac{\hat{g}_i \cdot \left(1 - \hat{g}_i\right)}{n_i - 1}, \tag{A9}$$

where n_i denotes the total number of observations at stimulus level x_i. Combining the last two results gives the standard error of $\hat{\mu}_{2AFC}$

$$\mathrm{SE}\left(\hat{\mu}_{2AFC}\right) = \sqrt{\sum_{i=1}^{k} \left(x_{i+1} - x_{i-1}\right)^2 \cdot \frac{\hat{g}_i \cdot \left(1 - \hat{g}_i\right)}{n_i - 1}}\ . \tag{A10}$$

APPENDIX B
The Area Above the Psychometric Function

If $x \geq 0$, then the area A bounded below by the psychometric function $G(x)$ and bounded above by the line $G(x) = 1$ is given by

$$A = \int_0^{\infty} \left[1 - G(x)\right] dx \tag{B1}$$

$$= \int_0^{\infty} \left[1 - (0.5 + 0.5 \cdot F(x))\right] dx \tag{B2}$$

$$= 0.5 \cdot \int_0^{\infty} \left[1 - F(x)\right] dx \tag{B3}$$

$$= \frac{\mu}{2}, \tag{B4}$$

since

$$\mu = \int_0^{\infty} \left[1 - F(x)\right] dx$$

must hold, where μ is the mean of $F(x)$ (see Feller, 1971, p. 150, Lemma 1). Therefore, the area A is equivalent to half the threshold value $\mu_{2\text{AFC}}$.

APPENDIX C
Redistributing Trials to Minimize the Standard Error of $\hat{\mu}_{\text{2AFC}}$

In this article, we have assumed that each stimulus level is tested in the same number of trials (i.e., $n_1 = n_2 = \cdots = n_k$). Although this procedure is simple and, thus, common practice in psychophysical research, it does not result in the most reliable estimate of $\mu_{2\text{AFC}}$. As is shown in this appendix, a more reliable estimate is obtained when the number of trials n_i is adjusted for each stimulus level x_i. In general, fewer trials should be employed for large stimulus levels at which performance is almost perfect, and more trials should be employed at small stimulus levels at which performance is close to chance.

For example, assume that the stimulus levels are $x = (2, 4, 6, 8, 10)$ and that the true associated probabilities of correct responses are $g = (.55, .62, .76, .88, .97)$. Assume also that the total number of trials is restricted to $n_1 + n_2 + \cdots + n_5 = N = 100$. Numerical computations based on the corresponding population version

$$\text{Var}\left(\hat{\mu}_{2\text{AFC}}\right) = 4 \cdot d^2 \cdot \sum_{i=1}^{k} \frac{g_i \cdot \left(1 - g_i\right)}{n_i} \tag{C1}$$

of Equation 10 reveal that the optimal numbers of trials per stimulus level are $n = (26, 26, 22, 17, 9)$, using the technique of Lagrange multipliers described next. The standard error for this optimally adjusted set of numbers of trials is 0.76. As is to be expected, this value is smaller than the $SE = 0.80$ obtained when equal numbers of trials are used at all stimulus levels, but the gain is only about 5% in this example. (One should note that the average estimated threshold does not depend on whether the numbers of trials are optimally adjusted or not. In either case, the average estimate would be 5.88 for the above example.)

In general, the optimal number of trials per stimulus level can be computed from Equation C1 subject to the constraint

$$N = \sum_{i}^{k} n_i$$

by the technique of Lagrange multipliers (e.g., Goldstein, Lay, & Schneider, 1987, pp. 367–377). According to this technique, one has to minimize the following Lagrange function:

$$H\left(n_1, n_2, \ldots, n_k, \lambda\right) = 4 \cdot d^2 \cdot \sum_{i=1}^{k} \frac{g_i \cdot \left(1 - g_i\right)}{n_i} + \lambda \cdot \left(N - \sum_{i=1}^{k} n_i\right) \tag{C2}$$

with respect to the variables $n_1, n_2, \ldots, n_k$, and λ. Computing the partial derivatives

$$\frac{\delta H}{\delta n_1}, \ldots, \frac{\delta H}{\delta n_k}$$

and

$$\frac{\delta H}{\delta \lambda},$$

APPENDIX C (Continued)

Table C1
Standard Error of the Estimate of $\hat{\mu}_{2AFC}$ as a Function of the Method of Analysis and the Choice of n_is

	Method of Analysis	
Choice of n_is	N	SK
Optimal n_is	0.641	0.737
Equal n_is	0.711	0.814
Reversed n_is	0.708	1.243

Note—N, normal; SK, Spearman–Kärber. See the text for simulation details.

setting these to zero, and solving the resulting system of equations for $n_1, \ldots, n_k$ yields the minimum of Equation C1. It can be shown that this minimum is given by

$$n_i = N \cdot \frac{\sqrt{g_i \cdot (1 - g_i)}}{\sum_{j=1}^{k} \sqrt{g_j \cdot (1 - g_j)}} \tag{C3}$$

for $i = 1, \ldots, k$. Therefore, the minimal standard error for $\hat{\mu}_{2AFC}$ is obtained when n_i is proportional to

$$\sqrt{g_i \cdot (1 - g_i)} \ .$$

Table C1 shows the results of simulations conducted to check how large an effect on standard error would be produced by optimizing the values of n_i. These simulations were conducted using the same procedures and simulation parameters as the simulations reported in the body of the article. We examined only the case in which there were five stimulus levels, a total of 120 trials divided across these five levels, and a true underlying normal distribution. Based on this true underlying distribution, the optimal values of n_i were found, using Equation C3, to be $n_i = (34, 34, 30, 17, 5)$. Simulations with these optimal n_is were compared against simulations with equal n_is (i.e., 24 per stimulus) and against simulations with the pattern of n_is reversed relative to the optimal pattern—that is, $n_i = (5, 17, 30, 34, 34)$. For each set of n_is, we generated 30,000 simulated data sets and computed $\hat{\mu}_{2AFC}$ from each data set, using both the Spearman–Kärber method and a probit analysis assuming an underlying normal distribution. The standard error of $\hat{\mu}_{2AFC}$ was estimated by the standard deviation of the obtained $\hat{\mu}_{2AFC}$ values across the 30,000 simulated data sets.

The results clearly show that the optimal choice of n_is produces the smallest standard error of estimation and that the effect is not trivially small. Relative to equal n_is, the optimal choice reduces standard error by approximately 10% in this example. In addition, the reversed pattern of n_is yields a standard error that is approximately 53% worse than equal n_is and 69% worse than the optimal choice. It is noteworthy that the choice of n_is that is optimal for the Spearman–Kärber method tends to yield the smallest standard error for the probit estimates, although the effect on probit estimates is much smaller. This is quite interesting given that the optimization method was derived on the basis of the Spearman–Kärber method.

The conclusion that optimal threshold estimation involves relatively more trials at the stimulus levels producing lower accuracies is slightly surprising because it appears in conflict with the idea that one should place trials at high probability correct (cf. Klein, 2001). This idea is supported, for example, by the work of Green (1990), who considered adaptive procedures for estimating threshold values in 2AFC tasks. Green showed that the smallest standard errors of estimate were obtained when the threshold was defined as a stimulus value yielding approximately 84%–94% correct, and on that basis he suggested that adaptive procedures should be designed to test the rather strong stimulus values associated with these high accuracy levels.

The discrepancy between the present conclusions and those of Green (1990) is apparent, not real, and in fact both conclusions stem from the same statistical properties of the estimators. The apparent discrepancy arises because of the differences in what is being estimated. Green considered the case of estimating a single target percentile, arbitrarily defined as the threshold. He showed that higher percentiles can be estimated more easily than lower ones and, on that basis, recommended estimating a rather high threshold. In contrast, we are here considering the problem of estimating the mean of the whole distribution with the Spearman–Kärber method. Estimating the mean of a distribution with the Spearman–Kärber method requires information about all of its percentiles (see Equation 7), not just about one of them. In that case, it is useful in terms of the overall estimate to have the most trials at the percentiles associated with the largest standard errors—that is, the ones close to .5 on the 2AFC psychometric function.

APPENDIX D
Effect of the Number of Stimulus Levels on the Standard Error of $\hat{\mu}_{2AFC}$

Assume that a researcher employs the set S_a of stimulus levels $S_a = (x_{1,a} < x_{1,a} < x_{2,a} < \cdots < x_{k,a})$ and that the spacing $d_a = x_{i,a} - x_{i-1,a}$ is constant for $i = 2, \ldots, k$. Furthermore, assume that the number of trials n_a per stimulus level is constant. Thus, the total number of trials is $N = k \cdot n_a$. Under this case, the expected variance for the sampling distribution of $\hat{\mu}_{2AFC}$ is

$$\mathrm{Var}\left(\hat{\mu}_{2AFC} \mid S_a\right) = \frac{4 \cdot d_a^2}{n_a} \sum_{i=1}^{k} p_{i,a} \cdot \left(1 - p_{i,a}\right). \tag{D1}$$

Now assume that the researcher doubles the number of stimulus levels and, thus, employs the set $S_b = (x_{1,b} < x_{2,b} < \cdots < x_{2 \cdot k,b})$ so that the extreme levels are equal for both sets—that is, $x_{1,a} = x_{2,b} = x_{\min}$ and $x_{k,a} = x_{2 \cdot k,b} = x_{\max}$. In other words, the two sets S_a and S_b cover the identical range on the x-axis. The spacing d_b between two consecutive stimulus levels is again constant for set S_b. Thus, the sampling variance of $\hat{\mu}_{2AFC}$ for this set is

$$\mathrm{Var}\left(\hat{\mu}_{2AFC} \mid S_b\right) = \frac{4 \cdot d_b^2}{n_b} \sum_{i=1}^{2 \cdot k} p_{i,b} \cdot \left(1 - p_{i,b}\right). \tag{D2}$$

It is possible to express $\mathrm{Var}(\hat{\mu}_{2AFC} \mid S_b)$ in terms of $\mathrm{Var}(\hat{\mu}_{2AFC} \mid S_a)$. First, note that the spacings d_a and d_b are given by

$$d_a = \frac{x_{\max} - x_{\min}}{k - 1}, \tag{D3}$$

$$d_b = \frac{x_{\max} - x_{\min}}{2 \cdot k - 1} \tag{D4}$$

and, therefore, the relation

$$d_b = d_a \cdot \frac{k - 1}{2 \cdot k - 1} \tag{D5}$$

must hold. Second, the relation

$$n_b = \frac{n_a}{2} \tag{D6}$$

must hold if the number of total trials is the same for both sets. Finally, the relation

$$\sum_{i=1}^{2 \cdot k} p_{i,b} \cdot \left(1 - p_{i,b}\right) = 2 \cdot \sum_{i=1}^{k} p_{i,a} \cdot \left(1 - p_{i,a}\right), \tag{D7}$$

should hold, at least approximately, since the sum on the left side contains twice as many terms as the sum on the right side.

Inserting Equations D5–D7 into Equation D2 yields

$$\mathrm{Var}\left(\hat{\mu}_{2AFC} \mid S_b\right) = 4 \cdot \left[\frac{k - 1}{2 \cdot k - 1}\right]^2 \cdot \mathrm{Var}\left(\hat{\mu}_{2AFC} \mid S_a\right). \tag{D8}$$

As an illustration, assume that a researcher doubles the number of stimulus levels from $k = 5$ to 10. In this case, the standard error would be reduced by a factor of $8/9 = 0.89$. An identical gain factor was obtained for the present simulations (i.e., 0.89) when the stimulus levels were increased from 5 to 10. As one might expect, a smaller gain factor is obtained if the number of levels is doubled from 10 to 20. In this case the standard error is reduced only by a factor of 0.95. One should bear in mind, however, that this conclusion rests on the validity of Equation D7 and, thus, can only be a rule of thumb.

(Manuscript received December 30, 2002;
revision accepted for publication September 12, 2003.)