Perception & Psychophysics
2001, 63 (8), 1330-1347

Fast and accurate measurement of taste and smell thresholds using a maximum-likelihood adaptive staircase procedure

MIRIAM R. LINSCHOTEN
University of Colorado Health Sciences Center, Denver, Colorado

LEWIS O. HARVEY, JR.
University of Colorado, Boulder, Colorado

and

PAMELA M. ELLER and BRUCE W. JAFEK
University of Colorado Health Sciences Center, Denver, Colorado

This paper evaluates the use of a maximum-likelihood adaptive staircase psychophysical procedure (ML-PEST), originally developed in vision and audition, for measuring detection thresholds in gustation and olfaction. The basis for the psychophysical measurement of thresholds with the ML-PEST procedure is developed. Then, two experiments and four simulations are reported. In the first experiment, ML-PEST was compared with the Wetherill and Levitt up–down staircase method and with the Cain ascending method of limits in the measurement of butyl alcohol thresholds. The four Monte Carlo simulations compared the three psychophysical procedures. In the second experiment, the test–retest reliability of ML-PEST for measuring NaCl and butyl alcohol thresholds was assessed. The results indicate that the ML-PEST method gives reliable and precise threshold measurements. Its ability to detect malingerers shows considerable promise. It is recommended for use in clinical testing.

Since the publication of *Elemente der Psychophysik* (Fechner, 1860), considerable effort has been made to improve the psychophysical methods used to measure sensory sensitivity (Guilford, 1936, 1954). We know that two psychological processes influence detection and discrimination performance: a sensory process and a decision process (Krantz, 1969; Swets, 1961; Swets, Tanner, & Birdsall, 1961). It is now possible to measure the sensitivity of the sensory process in a manner that is not influenced by the properties of the decision process. Such measures of sensitivity are said to be bias-free.

How one measures sensitivity depends on the model of the sensory process one adopts. Classical psychophysical methods assumed that the sensory process had a threshold, a stimulus value that must be exceeded in order for the stimulus to have any effect on the observer. These methods were therefore designed to measure the average value of this threshold. Sensitivity to a particular stimulus was represented by the probability, p, that the stimulus would exceed the sensory threshold. A stimulus that is equal to the average threshold value would have a p of .5. Other threshold models assumed that the sensory process had two or more sensory thresholds.

By the 1960s, however, it became clear that all models of the sensory process that assume one or more thresholds are invalid because they make predictions not supported by experimental data (Krantz, 1969; Swets, 1961, 1986a, 1986b, 1996). Sensitivity measures derived from these threshold models are contaminated by properties of the decision process.

Measures of sensitivity that are uncontaminated by decision-process properties (i.e., response bias) have been developed within the framework of signal detection theory (SDT). One almost bias-free measure of sensitivity is percent correct in the two-alternative forced-choice (2AFC) experimental paradigm (Green & Swets, 1966/1974; Macmillan & Creelman, 1991; Swets, 1986b). When percent correct (or probability of being correct) is plotted as a function of stimulus intensity, the resulting S-shaped curve forms what Urban (1908) first called a *psychometric function.* A typical 2AFC psychometric function formed by 20 stimulus intensities spanning a 3-log-unit range in 0.16-log-unit steps is illustrated in Figure 1. Each probability was computed from 1,000 presentations of the corresponding stimulus intensity.

Sensitivity is specified by the stimulus intensity required for a subject to reach a certain percent correct on

Parts of these data have been presented at the Eighteenth and Nineteenth Annual Meetings of the Association for Chemoreception Sciences, 1996 and 1997. The authors thank Stan Klein for his clarifying questions and helpful comments. The computer routines to carry out the computations are available from author L.O.H. at his Web site: http://psych.colorado.edu/~lharvey/. Correspondence should be addressed to M. R. Linschoten, Rocky Mountain Taste and Smell Center, UCHSC Box B-205, Denver, CO 80262 (e-mail: miriam.linschoten@uchsc.edu).

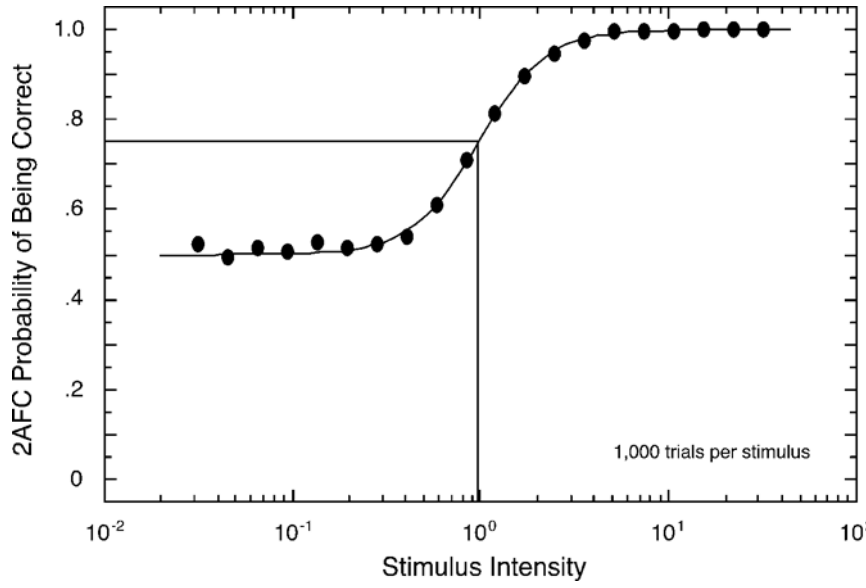

Figure 1. Typical S-shaped psychometric function for stimulus detection in a two-alternative forced-choice paradigm. The data were generated by a Monte Carlo logistic observer having an α of 1.0 stimulus intensity units, a β of 3.5, and a γ of 0.5. Each of the 20 stimulus intensities was presented 1,000 times. The solid line is the best-fitting logistic function fit to the data using a maximum-likelihood curve-fitting technique (Harvey, 1986, 1997).

the psychometric function. Typically, that performance point is halfway between chance and 100% correct performance. The 2AFC function has a halfway point of 75%. In the data shown in Figure 1, no stimulus gave a performance of exactly 75% correct, and this value must therefore be estimated by some suitable means. A logistic psychometric function (Berkson, 1951, 1953, 1955), shown as the smooth curve, was fit to the data using a maximum-likelihood technique (Harvey, 1986, 1997; Treutwein & Strasburger, 1999). The logistic function is given by

$$P(x) = \gamma + (1-\gamma)\cdot\left(\frac{1}{1+\left(\frac{x}{\alpha}\right)^{-\beta}}\right). \qquad (1)$$

The function is specified by three parameters: α (the stimulus at the halfway point), β (the steepness of the function), and γ (the probability of being correct by chance). The stimulus intensity corresponding to α is marked by the vertical line in Figure 1 and represents the stimulus that would achieve 75% correct in the 2AFC task (marked by the horizontal line). Even though the sensory process, as formulated in SDT, has no sensory threshold, the value of α is still widely referred to as the threshold. In this paper, the term *threshold* will be used to mean the stimulus value, α, that gives performance halfway from chance to 100%.

With the above framework in mind, it can be seen that the goal of all psychophysical methods for measuring thresholds and sensitivity, from the classical methods of Fechner (Guilford, 1936, 1954) to modern adaptive staircase methods (Cornsweet, 1962; Harvey, 1986, 1997; Treutwein & Strasburger, 1999; Watson & Pelli, 1983; Wetherill & Levitt, 1965), is to estimate the value of α. In vision and audition, it is practical to measure sensitivity with large numbers of trials using computer-generated stimuli. In the chemical senses, however, the physical presentation of the stimulus is not easily accomplished without human intervention. Furthermore, the longer recovery time in the chemical senses prevents rapid successive presentation of stimuli. These facts limit the total number of psychophysical trials that can be presented in a testing session before fatigue and/or boredom set in. We are therefore faced with the question: "How can we best estimate the threshold with a limited number of experimental trials?" A reasonable solution to this problem would be especially valuable for clinical testing of taste and smell function and for experiments in which sensitivity is repeatedly measured over a period of days or weeks.

A review of the taste and smell literature reveals that the two most widely used methods to measure detection sensitivity are the ascending method of limits for smell (Apter, Gent, & Frank, 1999; Cain, Gent, Catalanotto, & Goodspeed, 1983; Cometto-Muñiz & Cain, 1995; Deems et al., 1991; Duffy, Cain, & Ferris, 1999; Gagnon, Mergler, & Lapare, 1994; Lehrner, Brücke, Dal-Bianco, Gatterer, & Kryspin-Exner, 1997; Lehrner, Kryspin-Exner, & Vetter, 1995; Moll, Klimek, Eggers, & Mann, 1998; Murphy, Nordin, de Wijk, Cain, & Polich, 1994; Nordin et al., 1996; Pierce, Doty, & Amoore, 1996; Rosenblatt,

Olmstead, Iwamoto-Schaapp, & Jarvik, 1998) and the one-up–two-down variant of the Wetherill and Levitt staircase method (Wetherill & Levitt, 1965) for taste (Anliker, Bartoshuk, Ferris, & Hooks, 1991; Cowart, 1989; Cowart, Yokomukai, & Beauchamp, 1994; Drewnowski, Henderson, & Shore, 1997; Grushka, Sessle, & Howley, 1986; Murphy & Cain, 1986; Murphy et al., 1994; Stevens, 1995; Weiffenbach, Baum, & Burghauser, 1982; Weiffenbach, Schwartz, Atkinson, & Fox, 1995). The method of constant stimuli could be used to measure a complete psychometric function like the one shown in Figure 1 using a fixed number of stimuli presented numerous times in random order. Even though the shape of the function can be highly informative, this method is only rarely used in taste and smell research, most likely because of the large number of experimental trials required for stable estimates (Linschoten & Kroeze, 1991, 1992, 1994). A few experimenters have used an SDT paradigm (yes–no or rating scale judgments) to measure both sensory sensitivity and response bias (Doty, Snyder, Huggins, & Lowry, 1981; O'Mahony et al., 1979).

In audition and vision, adaptive psychophysical methods are widely employed, and the theoretical bases for them are well understood. To minimize the number of trials, the most efficient testing strategy is to compute an estimate of the threshold after each psychophysical trial and use a stimulus close to the threshold on the next trial (Taylor & Creelman, 1967). The estimation of the threshold is done by fitting a psychometric function to the data collected up to that point in the experiment and estimating the value of the threshold from the best-fitting function. This method is formally called a *maximum-likelihood adaptive staircase method* and is known by several different names: Best PEST (Pentland, 1980), QUEST (Watson & Pelli, 1983), and ML-PEST (Harvey, 1986, 1997). We will refer to it as the ML-PEST method (maximum-likelihood parameter estimation by sequential testing). The term PEST was originally coined by Taylor and Creelman (1967).

That good estimates of the threshold can be obtained from sparse data is shown in Figure 2, where psychometric functions for the same 20 stimuli as in Figure 1 but with 1,000, 100, 10, and 1 presentations per stimulus are shown. As the number of trials per stimulus intensity decreases, the empirical psychometric function becomes more and more irregular. The solid curve in each figure is the best-fitting logistic function (best value of α was found holding β and γ constant) obtained with a maximum-

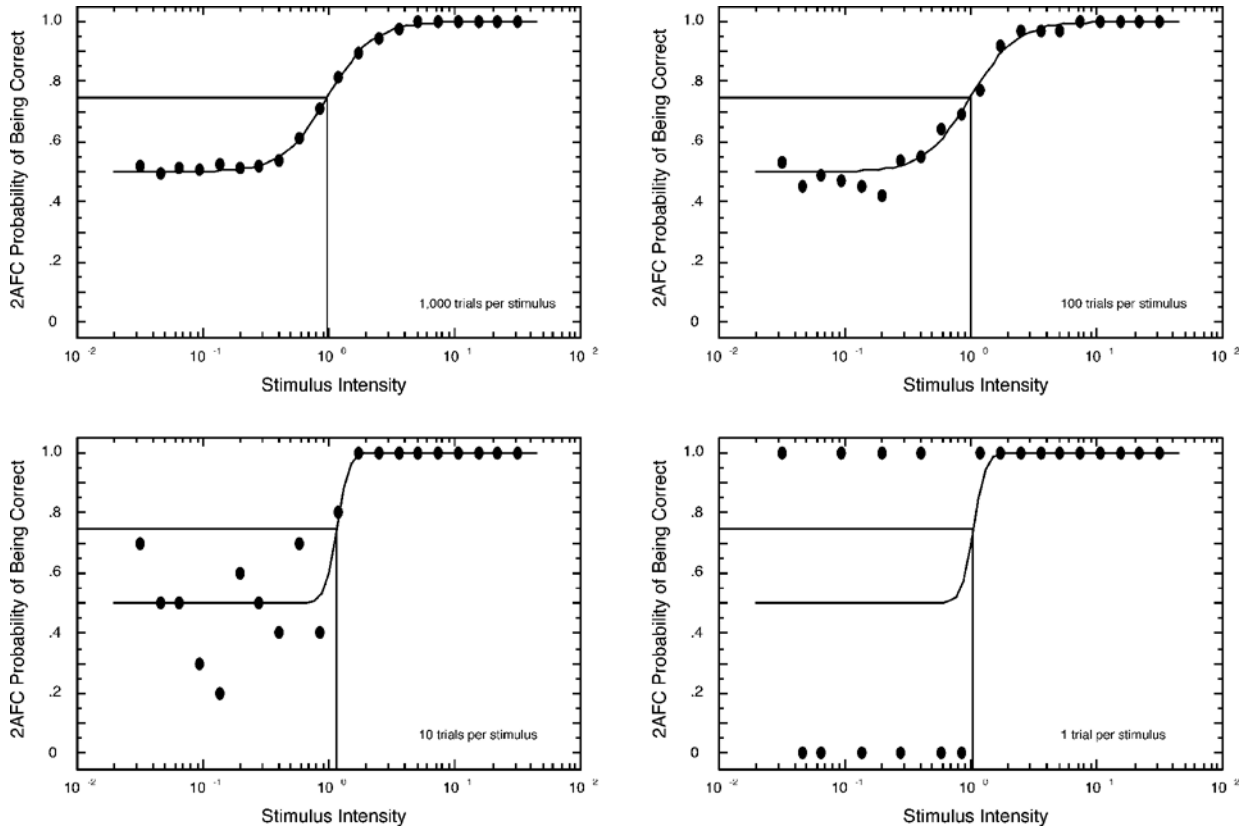

Figure 2. Typical S-shaped psychometric function for stimulus detection in a two-alternative forced-choice paradigm. The data were generated by a Monte Carlo logistic observer having an α of 1.0 stimulus intensity units, a β of 3.5, and a γ of 0.5. Each of the 20 stimulus intensities was presented 1,000, 100, 10, and 1 times. The solid line is the best-fitting logistic function fit to the data using a maximum-likelihood curve-fitting technique (Harvey, 1986, 1997).

likelihood fitting technique (Harvey, 1986, 1997). Even in the extreme case of 1 presentation per stimulus, the best-fitting logistic function gives a good estimate of the threshold.

Our goal in the present paper is to introduce the use of the maximum-likelihood adaptive staircase procedure for measuring sensitivity in the chemical senses. We will compare the efficiency and the accuracy of this method with those of other methods and present test–retest data as an assessment of reliability.

MAXIMUM-LIKELIHOOD ADAPTIVE STAIRCASE PSYCHOPHYSICAL METHOD

The principles of this method are described in a variety of sources (Hall, 1968, 1981; Harvey, 1986, 1997; Hays, 1963; Watson & Pelli, 1983), which provide more detail than is presented here. For this discussion, it is assumed that the psychometric function is logistic and that data are collected with a 2AFC experimental paradigm, although other functions and paradigms could certainly be used. During psychophysical testing, there are three major questions that are answered by this method after each trial: (1) What is the current estimate of the threshold? (2) What stimulus should be presented to the subject on the next trial? (3) Should the psychophysical trials be stopped? Although the answers are somewhat interrelated, it is useful to consider them separately.

Estimation of the Threshold

During psychophysical testing, the raw data collected on each trial are the stimulus concentration presented to the subject and whether or not the subject made a correct response. In order to estimate the threshold, a series of possible psychometric functions is considered. A likelihood function is constructed by computing the log likelihood for each of these possible psychometric functions, each one having the same value of β (steepness) and γ (the probability of being correct by chance), and each having a different value of α. The computational procedure is described by Harvey (1986, 1997). The smallest candidate α is chosen to lie below any possible real threshold or at least lower than the lowest physical stimulus. Successive values increase in 0.01-log-unit steps up to an appropriate maximum. In testing the detection of NaCl, for example, we consider 451 αs ranging from -4.50 log molar concentration (log M) to 0.00 log molar concentration in 0.01-log steps. Each candidate logistic function has a fixed β (3.5) and a fixed γ (0.5) for the 2AFC psychophysical paradigm. The β value is not critical for converging on the correct value of β. The value of 3.5 is close to actual psychometric functions for NaCl measured with the method of constant stimuli. The likelihood functions after having run 0, 5, 10, 15, 20, and 25 trials on a Monte Carlo observer with a true value of α of -2.125 log M (0.0075 M) are plotted in the upper panel of Figure 3. The best estimate of the threshold is that value of α having the maximum likelihood (i.e., the mean of the likelihood distribution). In this simulation, the best estimate of α was -2.110 log M after 25 trials: an error of 0.015 log M.

Choosing the Next Stimulus

At the beginning of an experiment, before any data have been collected, all candidate αs have equal likelihood of being the threshold. But, often, the experimenter has some prior idea of what the threshold value should be, especially if the subject has normal sensory function. We incorporate these prior expectations into the method by keeping track of a second likelihood function that starts with the prior likelihoods and adds the results of each trial to it. We use a Gaussian distribution for the prior likelihood function, with a standard deviation of 0.40, which is the value we found in our test–retest experiment to be reported below (Experiment 2). In the example, the initial estimate of the threshold was -1.875 log M. These posterior likelihood functions are plotted in the lower panel of Figure 3 after having run 0, 5, 10, and 15 trials on the same observer. The mode of this posterior likelihood function after each trial is used to choose the next stimulus.

In taste and smell, it is usually not feasible to maintain more than 20 or so different stimulus concentrations for presentation to an observer. The stimulus recommended by the maximum of the posterior likelihood function, therefore, may not actually be available to the experimenter to use on the next trial. There will be an actual stimulus that is lower than the recommendation and another stimulus that is higher than the recommendation. To choose between these two, we compute relative distance of the recommended stimulus between the two actual stimuli, and we compute the relative frequency of each of the two actual stimuli. We choose one randomly with a probability that favors the stimulus closer to the recommendation and the stimulus with the fewer number of trials. This computation is done in CStim->GetBestStimulusIndex() in the ML-PEST software package. Keeping the list of candidate αs separate from the list of available stimuli means that the threshold α can be computed to a precision that is not limited by the small number of actual stimuli used in the experiment.

For NaCl detection, we use 18 different concentrations ranging from -4.00 log M to 0.25 log M in 0.25-log-unit steps. The actual stimulus concentrations that were presented to the Monte Carlo observer are plotted in Figure 4 for each of the 25 trials. The solid circles represent trials on which the observer was correct; the open circles are incorrect trials. Because the observer was correct on the first three trials, the stimuli that were presented were successively weaker concentrations. On Trial 4, an error was made, and the stimulus on the following trial was the same concentration. Generally, but not always (because of the random process of choosing the next stimulus), the next stimulus will be weaker after a correct response and stronger after an incorrect response. The

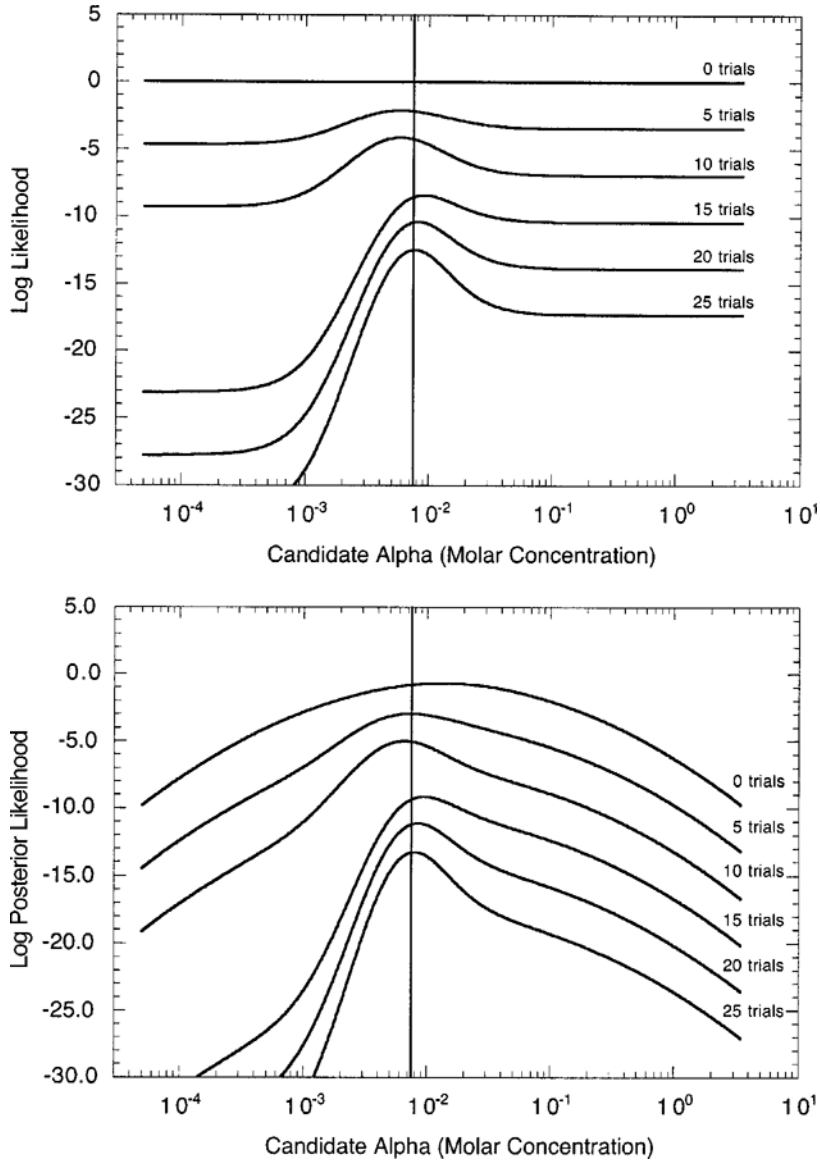

Figure 3. Log likelihood (upper panel) and log posterior likelihood (lower panel) as a function of candidate α following 0, 5, 10, 15, 20, and 25 Monte Carlo trials, with a logistic observer having a true α of 0.0075 M (−2.125 log M) detecting NaCl in a two-alternative forced-choice detection paradigm.

maximum-likelihood estimate of the threshold is plotted by the solid line in Figure 4. Until both correct and incorrect trials are collected, the estimate of the threshold will be at either the lower or the upper end of the range of candidate αs. In this example, the observer was correct on the first three trials and then made a mistake on Trial 4. The estimate of the threshold was extremely low (not visible in the figure) until Trial 4, when it jumped up to the region near the true value. As the trials continue, the estimated threshold fluctuates around the true value and eventually converges on it. After 25 trials, the estimated threshold was −2.11 log M: an error of −0.015 log units. This low error is achieved even though no actual stimulus corresponding to the threshold was available for testing.

When to Stop

Two stopping criteria can be used to terminate psychophysical testing: stop after a fixed number of trials or stop after the estimation of the threshold reaches a required level of precision. Precision is measured by the confidence interval, which is related to the width of the likelihood function. Notice in Figure 3 that the likelihood function becomes narrower as the number of trials

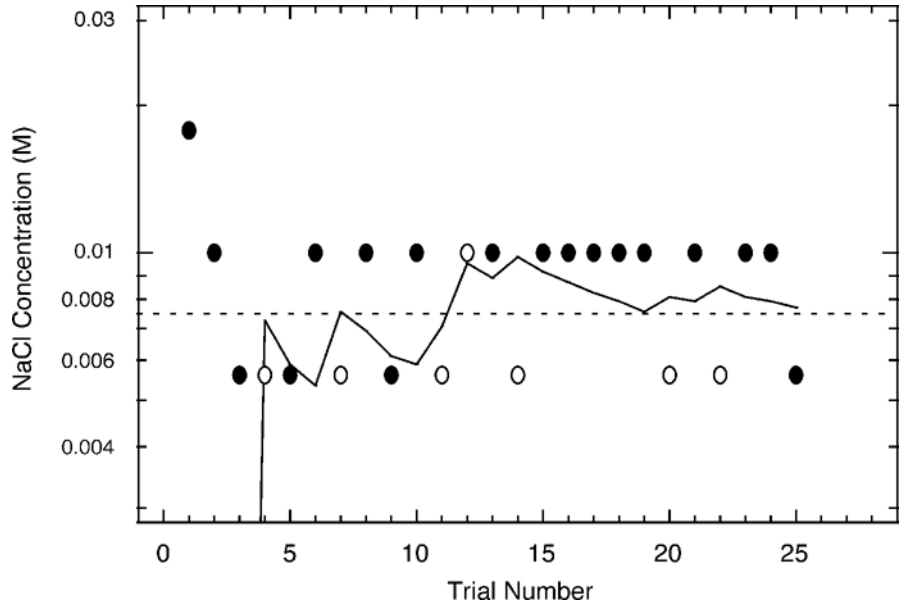

Figure 4. Stimulus presentation as a function of trial number in the two-alternative forced-choice Monte Carlo simulation. Trials on which the subject's response was correct are represented by the filled circles; trials on which the subject's response was incorrect are represented by the open circles. The resulting maximum-likelihood estimate of α is plotted by the solid line. The horizontal dotted line marks the true value of α.

increases. The confidence interval of the estimated threshold is directly related to the width of the likelihood distribution and is calculated using the likelihood ratio test and the χ^2 statistical distribution with 1 degree of freedom (Hays, 1963):

$$\chi^2 = -2.0 \cdot \log_e(\text{likelihoodratio}) = -2.0 \cdot \log_e\left(\frac{L_x}{L_{\max}}\right), \quad (2)$$

where L_x is the likelihood of a model whose α parameter is x and $L_{\max}$ is the likelihood of the best-fitting model. We search for the stimulus concentrations (x) above and below the threshold stimulus (the maximum-likelihood α) at which the log likelihood drops to $\chi^2/2$ below the maximum log likelihood value. To find the 95% confidence interval, the criterion value of χ^2 is 5.0239 (this value is obtained from standard tables of the χ^2 distribution found in all statistic text books: The one-tailed value of 5.0239 corresponds to a probability of .025; since the confidence interval is two-tailed, 5.0239 corresponds to a probability of .05).

In Figure 5, the same data shown in Figure 4 are plotted with an expanded vertical scale to show the estimate of the threshold along with the estimated 95% confidence interval. Notice that as the number of trials increases, the confidence interval of the threshold decreases. After 25 trials, the 95% confidence interval of the threshold extends 0.27 log units below the estimate of the threshold (indicated by the solid horizontal line, not by the filled circle, which represents the stimulus presented on that trial) and 0.37 log units above it, for a total confidence interval of 0.64 log units.

EXPERIMENT 1
Comparison of Three Psychophysical Methods

The goal of Experiment 1 was to compare the ML-PEST method with two commonly used methods with respect to their efficiency and their precision. Detection thresholds for butyl alcohol were measured using three methods: the ascending method of limits as described by Cain et al. (1983), the up–down staircase method (Wetherill & Levitt, 1965), and the ML-PEST maximum-likelihood staircase method as described above (Harvey, 1986, 1997).

Method

Subjects. Thirty-two (16 males and 16 females) healthy, nonsmoking subjects participated in the experiment. Their mean age was 30.4 years (± 8.1), ranging from 17 to 60 years. None reported having conditions that are normally associated with smell dysfunction.

Stimuli for ascending method of limits. For the ascending method of limits the procedure described by Cain et al. (1983) was used. Thirteen concentrations of butyl alcohol were prepared, with distilled water as diluent. The strongest concentration was 4%. Adjacent steps differed by a factor of 3.

Stimuli for staircase methods. The Wetherill and Levitt and ML-PEST methods used a series of 24 concentrations of butyl alcohol. The lowest 20 concentrations were a factor of 1.5 apart, with the strongest 0.1%. This series was augmented by the four strongest of the series for the ascending method, which were separated by a factor of 3.

Procedure. Thresholds were determined with a temporal 2AFC procedure. In each trial, a stimulus was paired with a blank, and the

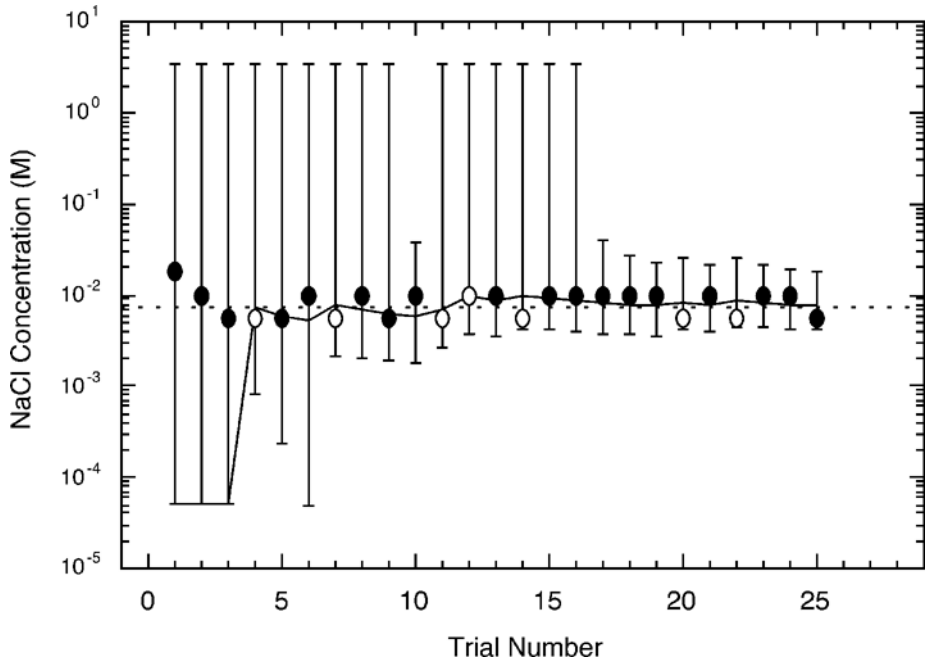

Figure 5. Stimulus presentation as a function of trial number in the two-alternative forced-choice Monte Carlo simulation. Trials on which the subject's response was correct are represented by the filled circles; trials on which the subject's response was incorrect are represented by the open circles. The resulting maximum-likelihood estimate of α is plotted by the solid line. The horizontal dotted line marks the true value of α. The vertical bars show the 95% confidence interval of each estimate of α.

subjects were asked to select the bottle containing the stimulus. The subjects were asked to give their best guess, even when they were certain that the two stimuli were identical. The stimulus and the blank were presented in 250-ml squeezable polyethylene bottles containing 60 ml of liquid. The subjects were instructed to keep one nostril closed while the other nostril was tested with three puffs out of each bottle. For each subject, four thresholds were obtained, two for each nostril. In randomized order, and alternating nostrils, one threshold was determined with the ascending method, one with the up–down method, and two with the adaptive maximum-likelihood method. Testing took about 45 min.

The definition of detection threshold differs in several aspects among the three methods. In the ascending method, testing started with the lowest concentration. A correct response led to the presentation of the same concentration on the next trial. When an incorrect response was given, the next higher concentration was used. Testing stopped when five correct answers had been given in a row. The threshold was defined as the last-used concentration (Cain et al., 1983) and, thus, corresponded to a detection probability of 1.0.

In the up–down staircase method, testing could start at any step in the concentration series. The first subject was started in the middle of the range, and each subsequent subject started where the former subject had ended. After one correct response, the same concentration was presented; if two correct responses were given, the next lower concentration was presented on the next trial. One incorrect response led to the presentation of a concentration one step higher. This decision rule is called the *one-up–two-down rule*. Although this decision rule is widely used, the actual threshold estimate depends more on the size of the stimulus step than on the particular up–down rule (Garcia-Perez, 1998). Testing stopped after five reversals, and the threshold was computed as the average of the stimulus concentrations at the last four reversals. This threshold corresponded to a detection probability of approximately .75.

In the maximum-likelihood adaptive staircase method, testing started in the middle of the stimulus range. After each response, the estimate of the threshold was calculated, and the next stimulus was a concentration close to that threshold. Testing stopped when the confidence interval of the estimated threshold reached a predefined criterion (0.8 log concentration units). This threshold corresponded to a detection probability of exactly .75. All these computations were carried out in real time on a Power Macintosh computer using a program, SensoryTester, written by author L.O.H.

Data analysis. The mean threshold (and standard error) produced by each method is presented in Table 1, labeled "Native Threshold." As noted above, each method produces a threshold estimate that corresponds to a different performance level of detection. To obtain threshold and confidence interval estimates for the ascending method of limits and the up–down staircase method that could be directly compared with those obtained by the ML-PEST method, a logistic psychometric function was fitted to the actual sequence of stimulus presentations and responses from these two methods. The values of β and γ were held constant at the same values used in the ML-PEST method. The maximum-likelihood threshold and the confidence interval were then computed in exactly the same manner as with the ML-PEST method. This fitting procedure will give a lower threshold for the ascending method of limits data than that reported by the method of limits procedure itself because the ascending method of limits looks for a threshold stimulus giving 100% correct, whereas the maximum-likelihood threshold corresponds to 75% correct. A repeated measures analysis of variance (ANOVA) was computed using StatView 5.01 (SAS Institute, 1998a, 1998b) on each of four dependent variables: native log threshold,

Table 1
Mean Performance (and Standard Error) of Three Psychophysical Methods With Butyl Alcohol in Experiment 1

	Native Threshold		Number of Trials		ML Logistic Threshold		Confidence Interval	
Method	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
ML-PEST								
Left	−3.41	0.16	15.13	0.65	−3.41	0.16	0.80	0.02
Right	−3.27	0.13	15.41	0.76	−3.27	0.13	0.84	0.02
Wetherill–Levitt	−3.30	0.10	17.44	0.72	−3.30	0.11	2.32	0.34
Ascending limits	−3.10	0.13	13.03	0.68	−3.48	0.14	1.14	0.02

number of trials to reach the native threshold, maximum-likelihood logistic threshold, and confidence interval of the maximum-likelihood threshold.

Results

The native threshold values, the mean numbers of trials needed to reach the stopping criterion, the maximum-likelihood logistic thresholds, and the mean confidence intervals of the maximum-likelihood thresholds are summarized in Table 1. A repeated measures ANOVA on the native thresholds showed no significant main effect of method [$F(3,31) = 1.61, p < .187$]. Subsequent statistical contrast comparisons, using the standard Huynh–Feldt ε to correct the p value and the degrees of freedom, revealed that the native threshold of the ascending method of limits was just significantly higher than the other methods [$F(1,31) = 3.92, p = .051$]. These higher thresholds are to be expected, given that the ascending method of limits computes the threshold to be the concentration corresponding to a detection probability of 1.0, whereas the other methods use a performance criterion of about .75.

A repeated measures ANOVA on the number of trials needed to reach the stopping criterion showed a significant main effect of method [$F(3,31) = 6.26, p < .0008$]. Statistical contrast comparisons showed that the ascending method required significantly fewer trials to determine threshold [$F(1,31) = 6.41, p < .014$] and that the up–down method required significantly more trials [$F(1,31) = 6.06, p < .017$] than the average of the two measurements with the maximum-likelihood method.

A repeated measures ANOVA on the maximum-likelihood logistic thresholds showed no significant main effect of method, and the thresholds computed from the trial-by-trial data from the ascending method of limits were not significantly different from the thresholds computed from the data collected with the other methods [$F(1,31) = 1.77, p = .19$]. This result is to be expected, since, presumably, the performance of each observer is driven by the same sensory process when tested by each of the three psychometric methods.

The mean confidence intervals for the thresholds (and standard errors) are shown in the last column of Table 1. The mean confidence interval of the maximum-likelihood method is smaller than those of the other two. Note that the standard error of the confidence interval reported for the ML-PEST method is very small (0.02). This is because the confidence interval was used as the stopping criterion: Trials were stopped when the confidence interval of α reached 0.8. A repeated measures ANOVA indicated that there was a significant main effect of method [$F(3,31) = 17.27, p < .0002$]. Contrasts tests revealed that the confidence intervals for the up–down method were significantly larger than those for both the maximum-likelihood method [$F(1,31) = 50.93, p < .0001$] and the ascending method [$F(1,31) = 23.72, p < .0016$].

Discussion

The native absolute threshold values obtained with the ascending method were higher than those obtained with the other two methods. One should remember that these values are not conventional thresholds, in the sense that they represent the concentration chosen correct five out of five times. It is to be expected that this value will be a higher concentration than thresholds based on a lower percent correct. Furthermore, this method, at least as described by Cain et al. (1983), has considerably larger steps between adjacent concentrations, and only actual stimulus concentrations can be threshold values.

The ascending method needed the fewest number of trials (~13) relative to either the maximum-likelihood method (~15) or the up–down staircase method (~17). The clear superiority of the maximum-likelihood method, however, came with the high precision of the measured thresholds, as indicated by the small confidence intervals. The confidence intervals of the up–down staircase were almost three times larger than those given by the maximum-likelihood method.

MONTE CARLO SIMULATIONS

When measuring the threshold of an actual subject, it is not possible to know what the error is between the measured value and the "true" value because the true value is not known. In order to evaluate the accuracy of the three psychophysical methods, we resorted to Monte Carlo evaluations of simulated observers having a predetermined threshold value. We used each of the three psychophysical methods to measure the threshold of each simulated observer and compared the result with the true value.

Monte Carlo simulations were run for each of the three psychophysical procedures using the same rules and stimulus series as were used in the threshold mea-

Table 2
Results of the Four Monte Carlo Simulations

Method	Median Error	Median Number of Trials
Simulation 1		
Maximum-likelihood	−0.027	16
Wetherill–Levitt	−0.045	17
Ascending limits	0.331	14
Simulation 2		
Maximum-likelihood	0.094	17
Wetherill–Levitt	0.101	17
Ascending limits	0.419	14
Simulation 3		
Maximum-likelihood	−0.03	16
Wetherill–Levitt	3.60	16
Ascending limits	3.54	16
Simulation 4		
Maximum-likelihood	2.75	115
Wetherill–Levitt	0.93	21
Ascending limits	3.36	22

surements for butyl alcohol with the real subjects reported above in Experiment 1. A population of 10,000 simulated subjects was defined whose thresholds followed a Gaussian distribution with a mean threshold of −3.00 log percent concentration and a standard deviation of 0.4 log units. These values are representative of the distribution of real subjects' detection thresholds for butyl alcohol, as found in Experiment 2 below. The predetermined threshold for each of the 10,000 simulated observers was drawn from the above distribution. Threshold values higher than the highest stimulus concentration and lower than the lowest stimulus concentration were not used.

Simulation 1

On each trial, the simulated subject was presented with a stimulus. Each subject's detection behavior was driven by a logistic psychometric function with a steepness of 3.5 and a chance level performance of 0.5. The probability of making the correct response on a trial was determined by the probability predicted by the logistic function for the stimulus and for that subject and a uniform random number between 0 and 1. If the random number was less than the logistic function probability for that stimulus, the response was scored as correct; otherwise, it was scored as wrong. The starting stimulus for the up–down staircase and maximum-likelihood methods was chosen randomly between the two stimuli that bracketed the mean threshold of −3.0 log concentration. These stimuli were −3.113 and −2.937 log concentration. The ascending method started with the lowest stimulus in its series, −5.130 log percent concentration.

An ascending limits run was considered a failure if the procedure called for a stimulus stronger than the highest stimulus concentration. A Wetherill–Levitt staircase run was considered a failure if, during the trials, a stimulus stronger than the highest stimulus concentration or weaker than the lowest stimulus concentration was called for. In the present simulation, the Wetherill–Levitt staircase failed 66 times. All of these failures were caused by the procedure's calling for a lower stimulus concentration than the lowest available. An ML-PEST run would have been considered a failure if the final estimate of the threshold were more than half the stopping confidence interval below the lowest candidate or above the highest candidate α. There were no such failures. The failed runs are excluded from the data analyzed below.

The distribution of errors for each method (estimated threshold minus true threshold) for the 10,000 subjects is shown in the left side of Figure 6. The distribution of the number of trials to reach stopping criterion for the 10,000 subjects is shown on the right side of Figure 6. The maximum-likelihood method was more accurate than the Wetherill–Levitt staircase, and both of these methods were superior to the ascending method of limits. The median error and the number of trials required for each method are shown in Table 2. The median error for the maximum-likelihood method was almost half as large as that for the up–down staircase, although both errors were quite small.

The error on each of the 10,000 subjects is plotted in Figure 7 as a function of the number of trials for that run. The range of errors became smaller as the number of trials increased for both the Wetherill–Levitt staircase method and the ML-PEST method, and the median error stayed close to zero. The ascending staircase method showed quite different behavior. The direction of the error depended on the number of trials before the stopping criterion of five correct in a row had been reached. The estimated thresholds were too low when the number of trials was less than about 15, whereas the estimates of threshold were too high when the number of trials was more than 15.

Simulation 2

Because the Wetherill–Levitt staircase method and the ML-PEST method generally required different numbers of trials before the stopping criterion was reached on each run, comparison of the accuracy is not straightforward. We repeated the Monte Carlo simulations so that the staircase method and the ML-PEST method used the same number of trials on each run. First, the Wetherill–Levitt staircase was run until the stopping criterion was reached. Then, the ML-PEST method was run with the same number of trials on that observer using the same starting stimulus.

The results are presented in Table 2. When the ML-PEST method had the same number of trials as the up–down staircase, the ML-PEST method was slightly more accurate. As in Simulation 1, both of these methods were superior in accuracy to the ascending staircase. There were 60 failures out of the 10,000 cases where the up–down staircase called for a stimulus that was lower than the lowest available concentration. Because the ML-PEST method was not permitted to run as many trials as it would normally require, this method failed on 30 out of the 10,000 cases.

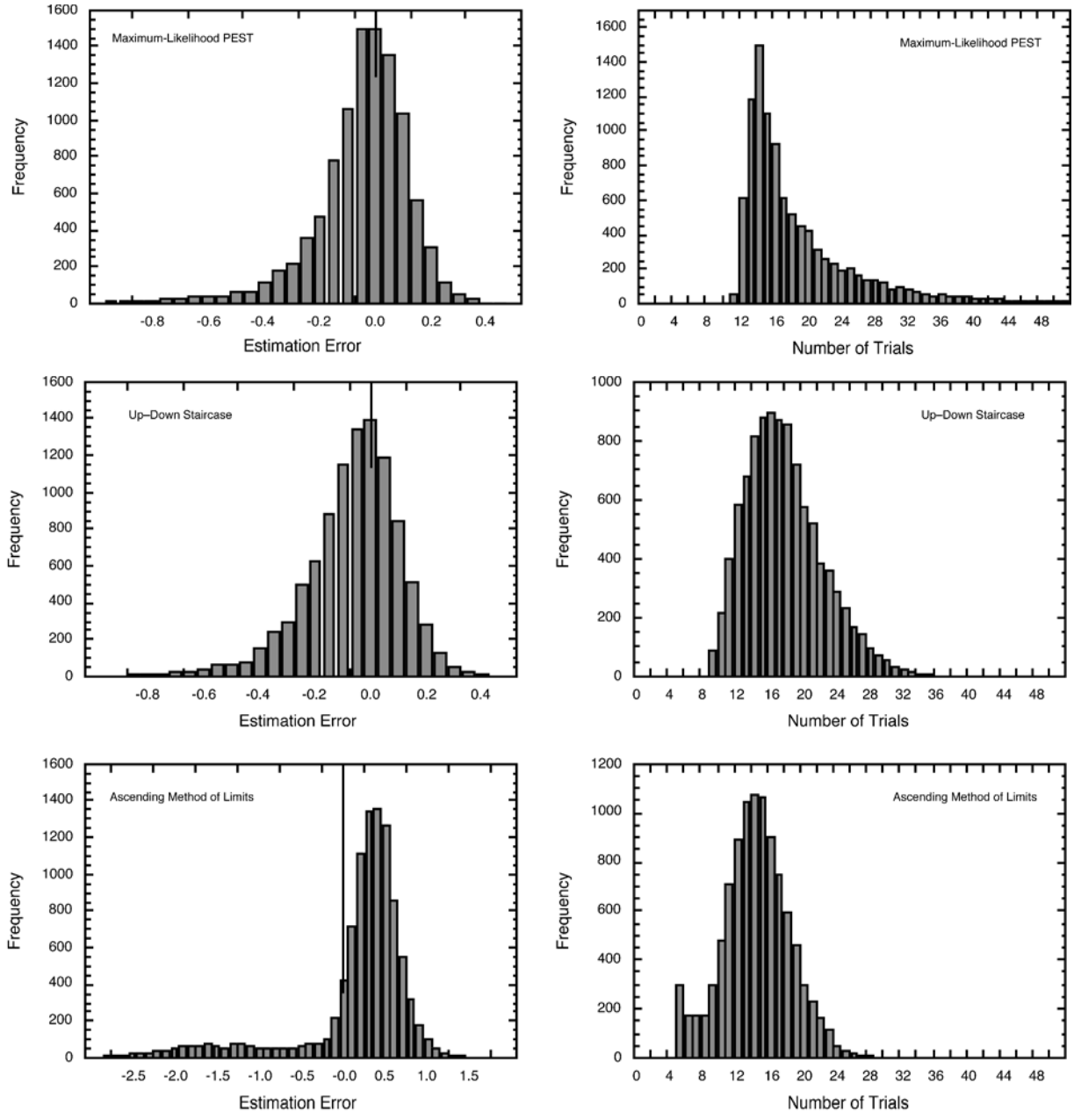

Figure 6. Distribution of errors (left) and number of trials (right) for three psychophysical methods: maximum-likelihood adaptive staircase (upper panel), Wetherill and Levitt up–down staircase (middle panel), and ascending method of limits (lower panel). The vertical line in the left panels marks the zero-error bin of each histogram. See text for details. The abscissa scale for the upper two left panels covers a range of 1.5 log units and is 5 log units for the lower left histogram.

Simulation 3

An important ability of sensory testing is to detect people who deliberately try to appear to have diminished sensitivity. We therefore repeated the Monte Carlo simulations of 10,000 subjects using the same starting conditions and the same test conditions as those in Simulation 1, but each simulated subject generated "untruthful" responses (i.e., responses just the opposite of that called for by the subject's psychometric function). In the ML-PEST computer program, two likelihood functions are maintained: one based on the hypothesis that the subject is responding truthfully and the other based on the hypothesis that the subject is deliberately choosing the wrong response. The program explicitly tests the hypothesis that a subject is deliberately choosing the wrong response by comparing the maxima of the two likelihood distributions and, therefore, should be able to measure a sensory threshold as effectively as when a subject is

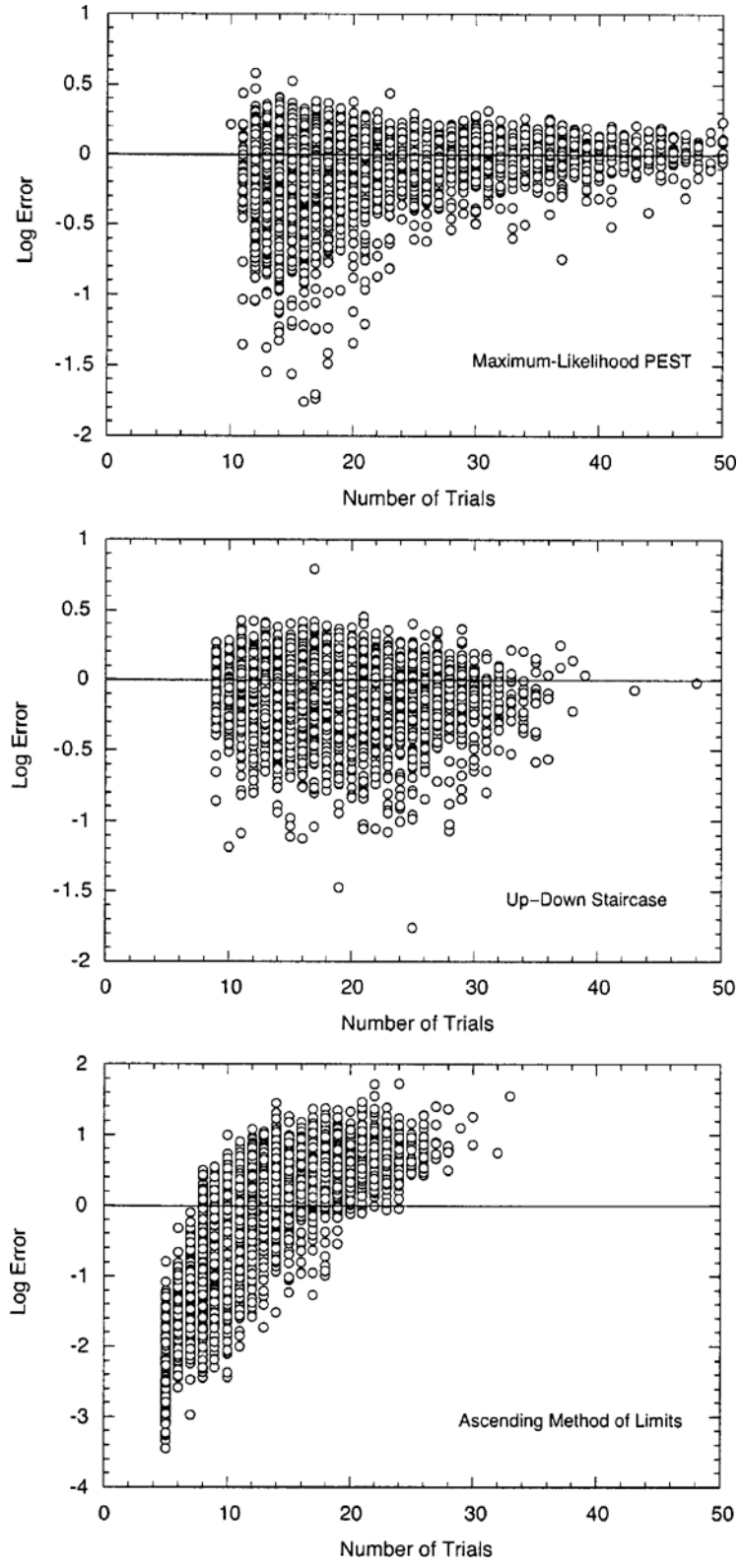

Figure 7. Error as a function of number of trials for three psychophysical methods: ML-PEST (upper panel), up–down staircase (middle panel), and ascending method of limits (lower panel).

telling the truth. The other two methods do not have an explicit way of dealing with this situation. It is of interest to learn how all three methods behave under these circumstances.

The results of this simulation are presented in Table 2. The ML-PEST method successfully measured the threshold and explicitly detected the lying condition in all 10,000 cases and was therefore able to measure the threshold of the underlying psychometric function with a small median error. The up–down staircase method and the ascending method of limits reached the highest concentration and tried to go higher, resulting in 9,824 failures for the staircase method and 8,648 failures for the method of limits. If we consider a failure as a successful detection of the lying behavior, the up–down staircase method detected 98% of the cases, and the ascending method of limits detected 86% of the cases. These two methods required about the same number of trials (see Table 2).

Simulation 4

It is important in clinical testing that people with elevated thresholds be easily detected. One of the subjects in Experiment 1 first had smell tested by the up–down method and exhibited a normal threshold. Later, the same nostril was tested using the maximum-likelihood method, and no threshold could be measured because, as it turned out, he was anosmic on that side. The fact that the up–down method did not detect the anosmia is of concern.

We therefore repeated the Monte Carlo simulations of 10,000 subjects using the same starting conditions and the same test conditions as those in Simulation 1, with one change: Each simulated subject generated a random response, as would be the case if a person had no sensory sensitivity. The ideal behavior of a testing method would be to indicate that the threshold was at the highest value that is possible with that method or to have some other explicit way of indicating that the subject is responding randomly.

None of the three methods performed very well in detecting the random responding. The median errors and number of trials are given in Table 2. The up–down staircase method did the poorest because it gives a median error of .93. This rather low error occurs because, on 87% of the cases, the stopping criterion of 5 reversals was satisfied before the method reached the upper limit of the stimulus concentrations. The ML-PEST method failed on 11% of the cases, but, overall, the thresholds were estimated to be very high. Unfortunately, it required a great number of trials to achieve this performance. The ascending method of limits reached the upper limit of stimulus concentration on 66% of the cases and also gave the highest values for the threshold.

Since, in Monte Carlo simulations, the distribution of the actual thresholds is known, we can apply an SDT analysis (Green & Swets, 1966/1974; Macmillan & Creelman, 1991) to compare the distribution of estimated thresholds made by each method under random responding with the true distribution. Using the means and standard deviations of the above distributions, we computed the signal detection measures of sensitivity d_a and accuracy, the area under the ROC curve, A_z (Harvey, 1992; Simpson & Fitter, 1973). The ML-PEST method had the highest sensitivity and accuracy ($d_a = 2.27, A_z = .95$). The ascending method was next best ($d_a = 1.99, A_z = .92$), and the up–down staircase method was the least sensitive ($d_a = 1.50, A_z = .86$).

Discussion

The results of the simulations lead to the conclusion that both the up–down staircase method and the maximum-likelihood method have considerable strengths for use in measuring taste and smell thresholds. The ascending method of limits should not be used if one is interested in obtaining an accurate estimate of the threshold. These threshold estimates covary with the number of trials taken to stop, as is illustrated in the lower panel of Figure 7. The simulation results offer an explanation for the published finding that ascending limit measurement of olfactory thresholds are less reliable than staircase measured thresholds (Doty, McKeown, Lee, & Shaman, 1995).

The up–down staircase method gives an average error twice as large as that of the ML-PEST method, although both values are small and, therefore, similar to each other. The up–down staircase method failed by calling for a stimulus too low on 66 out of 10,000 cases, whereas the ML-PEST method had no such failures. Although the median number of trials gives the edge to the ML-PEST method, on some occasions the ML-PEST method required over 50 trials to reach its stopping criterion.

A strength of the ML-PEST method is its ability to consider alternate hypotheses about the response strategy of the observer. The ML-PEST method was able to correctly detect all 10,000 dishonest responders. We are currently preparing a paper that will go into considerable detail on how to distinguish malingerers from true anosmics using a variety of methods. Another strength is that there is virtually no bias introduced into threshold measurements by having the ML-PEST steepness different from the "true" value set in the Monte Carlo observer (Treutwein & Strasburger, 1999). Data generated from a Monte Carlo Weibull function observer can be well fit by virtually any *s*-shaped psychometric function, although the Weibull will usually fit slightly better. We therefore recommend the ML-PEST method for testing taste and smell.

EXPERIMENT 2
Test–Retest Reliability

In order to assess the stability of the measures obtained with the adaptive maximum-likelihood procedure, we tested the subjects on four occasions.

Method

Subjects. Twenty-seven nonsmoking, healthy subjects participated in the experiment (mean age = 31.6 ± 8.7 years, range =

22–57 years). The 14 women and 13 men were divided into two groups. Group 1 ("consecutive," 7 females and 7 males) was tested on 4 consecutive days, and Group 2 ("distributed," 7 females and 6 males) was tested on Days 1, 2, 8, and 22.

Stimuli. Smell thresholds for butyl alcohol were assessed using the same range of concentrations as were used with the maximum-likelihood method in Experiment 1. Taste thresholds were measured using 20 concentrations of NaCl in quarter-log steps. The highest concentration was 1.7 M. The same maximum-likelihood adaptive method as described in Experiment 1 was used in this experiment.

Procedure. Thresholds were determined with a temporal 2AFC procedure. Taste sensitivity was measured with a whole-mouth sip-and-spit procedure. The stimuli and the blanks were presented in 30-ml plastic medicine cups, containing 5 ml of liquid. The subjects were requested to take the first sample in their mouths, swish it around, and spit it out, then rinse with distilled water, and do the same with the second sample. They were asked to choose the sample that had a taste different from water. In each of the four sessions, taste thresholds were determined first and smell thresholds last. All four smell thresholds were determined in the preferred nostril. The subjects were tested at the same time of each day.

Results

Because the measurement and specification of test–retest reliability is a rather complex topic (Linn, 1989), we have chosen four different ways to demonstrate the reliability of the maximum-likelihood method. For the first method, the four threshold measurements for each subject are plotted in Figure 8. The thick line in each panel is the mean threshold. It is clear from examination of Figure 8 that most individual thresholds were stable over 4 days (left column of panels) and over 22 days (right column of panels), with a few subjects showing shifts of more than a 1 log unit. The mean thresholds for the two groups are shown in Table 3. A repeated measures ANOVA confirmed the stability of the thresholds: The effect of testing day was not significant.

A second approach is to compute the correlations among the four threshold measurements (see, e.g., Mattes, 1988). Pearson correlation coefficients between thresholds obtained on different days are shown in Table 4. The correlations were generally higher for pairs of thresholds that were measured close together in time. As Mattes pointed out, correlations are highly influenced by small changes in ordinal ranking of the subjects. The fact that the pairwise correlations became smaller as time separation between them increased is shown in Figure 9. The

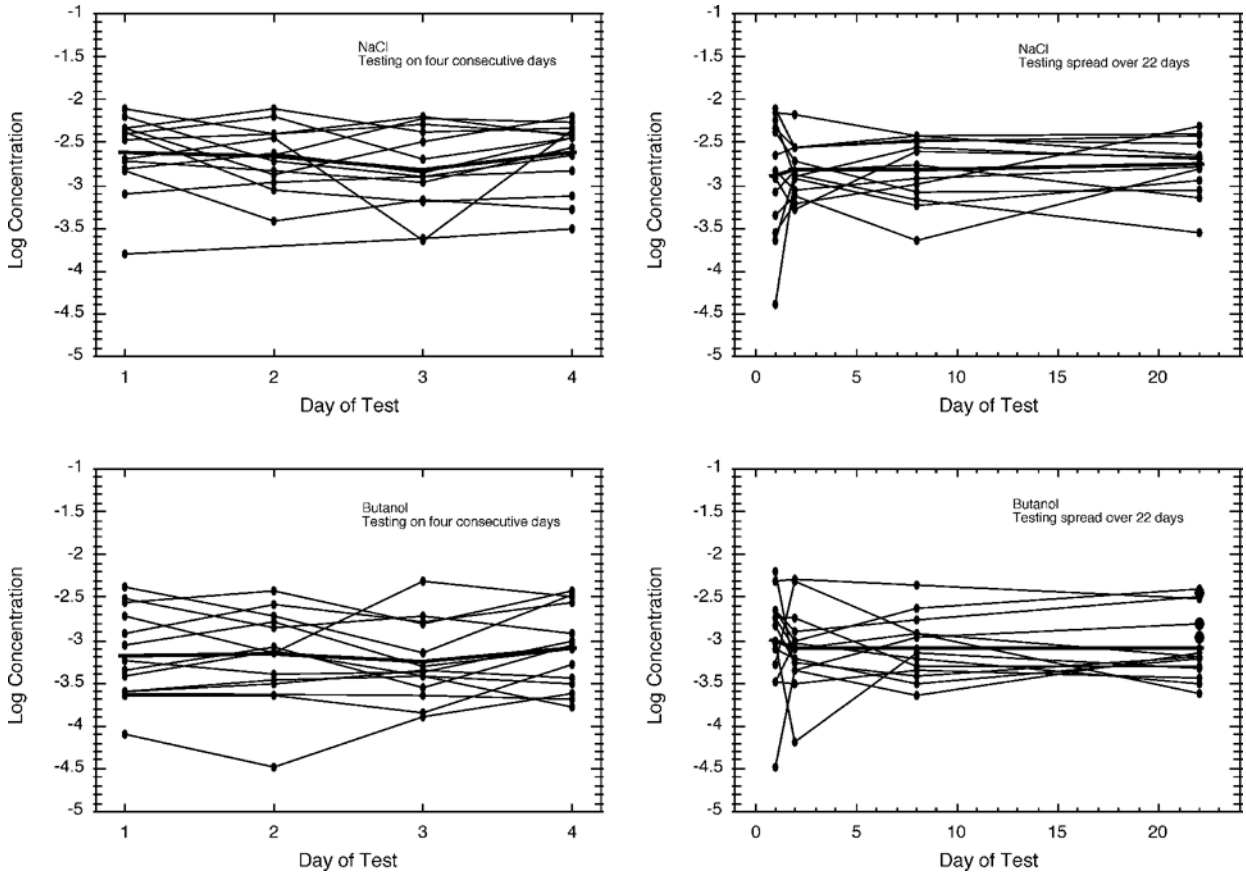

Figure 8. Log threshold concentration as a function of test day for NaCl (upper row) and butanol (lower row) measured on 4 successive days (left column) and over 22 days (right column). The thin lines connect the four thresholds of individual subjects. The thick line is the mean threshold concentration of all subjects.

Table 3
Mean NaCl and Butyl Alcohol Thresholds (and Standard Errors) From Experiment 2

	Day 1		Day 2		Day 3		Day 4	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
NaCl								
Consecutive (Group 1)	−2.62	0.13	−2.67	0.11	−2.82	0.13	−2.63	0.11
Distributed (Group 2)	−2.90	0.19	−2.83	0.09	−2.83	0.11	−2.77	0.10
Butyl Alcohol								
Consecutive (Group 1)	−3.19	0.14	−3.17	0.15	−3.25	0.12	−3.10	0.13
Distributed (Group 2)	−3.00	0.16	−3.10	0.14	−3.09	0.10	−3.09	0.11

solid line is the least squares regression line. It indicates that the correlation between two thresholds diminished as the time between them increased. The slope of the regression line becomes even steeper if the data for 1-day separation are removed from the regression, indicating that the decrease in correlation started with the 2-day separation condition.

The third method to assess test–retest reliability is based on the differences between successive threshold measurements. Each threshold was subtracted from the previous threshold: The four threshold measurements resulted in three difference measurements. The distributions of these differences for NaCl and for butyl alcohol are plotted on the left side of Figure 10. If the measured thresholds do not systematically increase or decrease over time, we expect that the mean difference would be zero. The mean of the distribution for NaCl (upper panel of Figure 10) was −0.018, and for butyl alcohol (lower panel) was 0.002. Neither of these was significantly different from zero. The standard deviations of these distributions were 0.413 log units for NaCl and 0.422 log units for butyl alcohol.

A fourth way to examine test–retest reliability is by means of the spread of the four threshold measures for each subject. A standard deviation of 0.0 would mean that all four thresholds were exactly the same value. The standard deviation of the four measurements was computed for each subject. The distribution of these standard deviations for NaCl (upper panel) and butyl alcohol (lower panel) are plotted on the right side of Figure 10. The median standard deviation was about 0.15 log units for both NaCl and butyl alcohol. This good test–retest reliability is consistent with the other three methods of assessing reliability reported above.

Discussion

The maximum-likelihood method gives estimates of thresholds that are stable within individuals over time and that are responsive to individual differences. By all four indicators, the test–retest reliability of both NaCl and butyl alcohol thresholds measured by the maximum-likelihood method was very good.

GENERAL DISCUSSION

The two experiments and the Monte Carlo simulations reported here demonstrate that the maximum-likelihood method is a viable alternative to other methods currently in use to measure taste and smell detection thresholds. Because the method attempts to use stimuli that are optimal (close to the predicted threshold), the confidence intervals of the measured thresholds are considerably smaller than those achieved by the other two methods (Experiment 1).

One of the strengths of the maximum-likelihood method is that it provides an explicit way of estimating the confidence interval of the threshold, a facility not provided by the other two methods. The experimenter is free to establish a stopping confidence interval to suit his/her needs. If greater precision is desired, it can be achieved at the expense of running more trials. This explicit tradeoff between precision and number of trials is not possible with the other two psychophysical methods. One could, of course, use a hybrid combination of methods: an up–down staircase method for generating the next stimulus to use combined with a maximum-likelihood method for estimating the threshold and its confidence interval. In effect, we used this approach after the data from Experiment 1 were collected by reanalyzing them

Table 4
Pearson Correlation Coefficients Between Thresholds

	Paired Sessions					
	1–2	1–3	1–4	2–3	2–4	3–4
NaCl						
Consecutive (Group 1)	.46	.66*	.66*	.44	.75**	.66**
Distributed (Group 2)	.58*	.72**	.40	.55*	.20	.51
Butyl alcohol						
Consecutive (Group 1)	.86***	.83***	.86***	.67**	.71**	.76**
Distributed (Group 2)	.26	.68**	.56*	.51*	.21	.65*

*$p < .05$. **$p < .01$. ***$p < .001$.

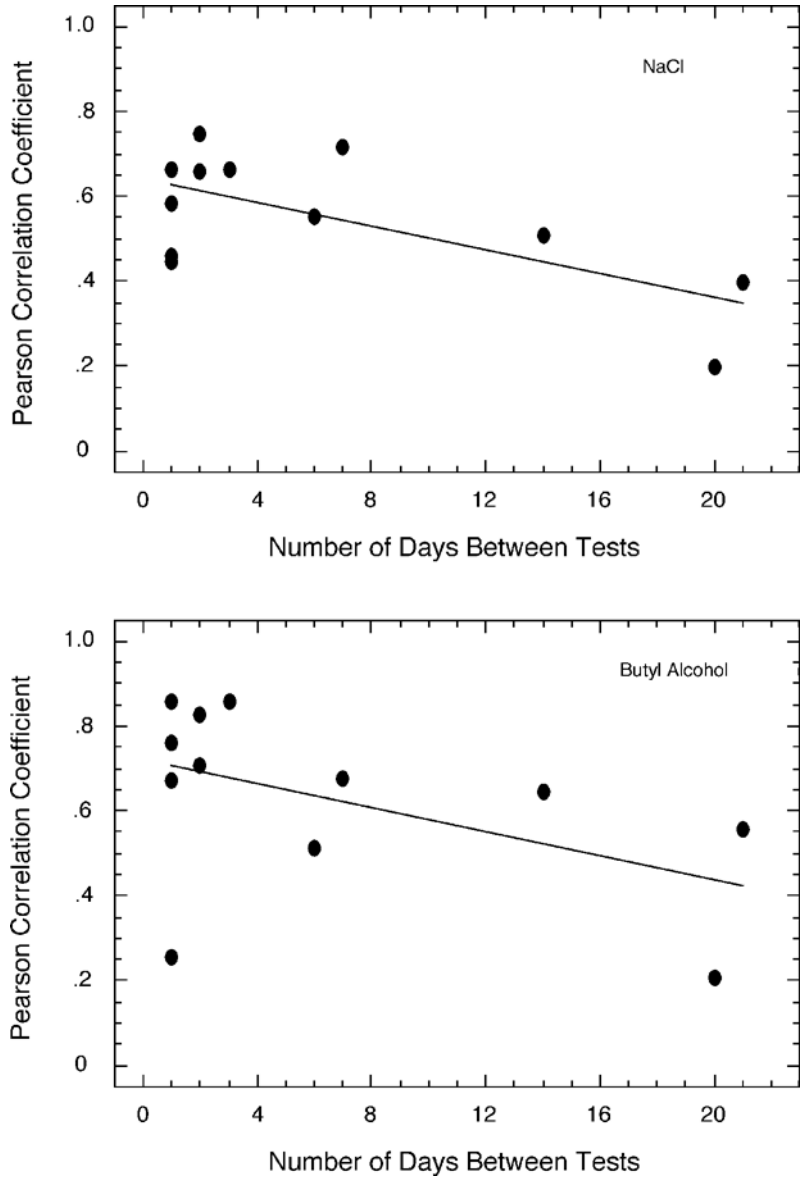

Figure 9. Test–retest reliability as indicated by correlation coefficients between all pairs of measurements plotted as a function of the number of days separating them. The solid line is the least squares linear regression through the data.

with the maximum-likelihood technique, in order to assess the confidence interval of the resulting threshold.

The Monte Carlo simulations allow the analysis of the ideal behavior of the three psychophysical methods. One major conclusion is that the ascending staircase method gives threshold measurements that are severely biased. This bias is a function of the number of trials carried out before the stopping criterion of 5 correct in a row with one stimulus. When the number of trials was small (less than approximately 15), the threshold estimate was too low; when the number of trials was greater than 15, the threshold estimate was too high (see Figure 7). This characteristic is due to the fact that the stimulus concentration used can only increase as trials progress. The maximum-likelihood method has the least amount of bias, although it is always slightly negative (i.e., the threshold estimate is, on the average, slightly lower, by 0.02 log units, than the true value).

In clinical testing, especially for cases that involve the legal system, it is important to be able to detect malingerers. One of the strengths of the maximum-likelihood method is that it explicitly tests hypotheses. Normally, the hypothesis is that the subject is responding truthfully and that the threshold has a certain value. In the present

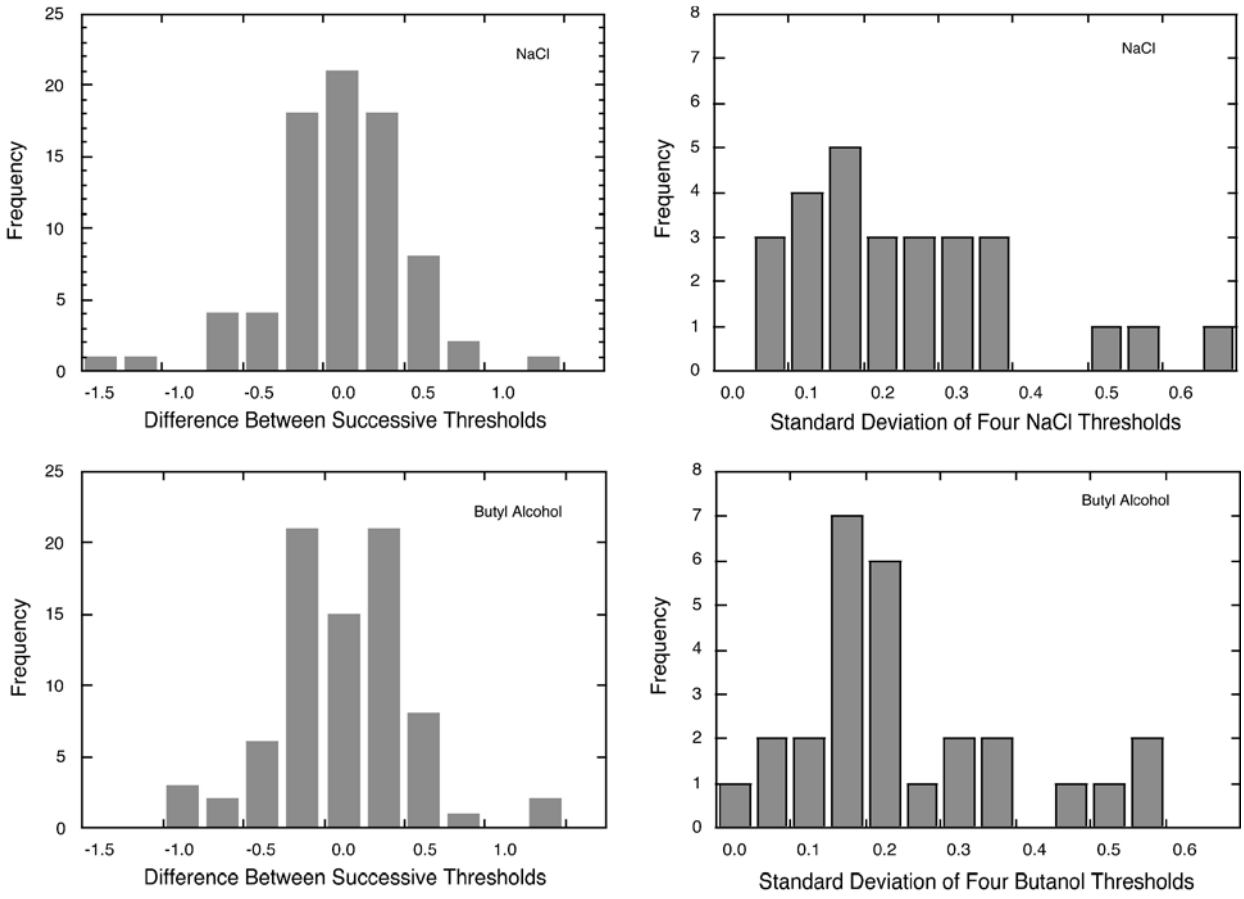

Figure 10. Distribution of differences between successive thresholds (left) and of the standard deviation of the four threshold measurements for the 27 subjects (right) for NaCl (upper panel) and butyl alcohol (lower panel).

implementation of the maximum-likelihood method, the computer routines maintain three hypotheses about the response strategy of the subject: that the subject is responding truthfully, that the subject is lying on each trial (i.e., always choosing what he/she believes is the wrong alternative in the 2AFC procedure), and that the subject is responding randomly (i.e., not making use of any sensory information). The ability to detect a malingerer—one who has sensory sensitivity but tries to respond randomly—is based on the fact that it is difficult for humans to generate truly random sequences (Bar-Hillel & Wagenaar, 1993). A subject who can taste or smell the stimulus and therefore knows in which interval it occurs tends to respond more wrongly than they would by chance alone. In this case, the hypothesis that the subject is lying will have a higher likelihood than the hypothesis that the subject is telling the truth. We are currently working to refine this ability of the maximum-likelihood method, because it holds great promise. The other two psychophysical procedures do not easily distinguish between a malingerer and a person with no sensory sensitivity and have no explicit way to detect malingerers.

Finally, in Experiment 2, we demonstrated that the maximum-likelihood method gives estimates of thresholds that are stable over time. This stability is actually based on two aspects of our testing procedure. The maximum-likelihood method gives the smallest confidence intervals of the three methods tested and thus contributes the least amount of random or systematic error variance to the measured thresholds. The second source of reliability is the experimental paradigm: The 2AFC procedure minimizes the effects of response bias (e.g., Green & Swets, 1966/1974) and thus minimizes error variance in the threshold measurements that would be introduced by variance in response bias. Actually, it would be desirable to use more than just two forced-choice alternatives, because the ML-PEST methods converges more rapidly on the threshold value with more alternatives (Manny & Klein, 1985; Schlauch & Rose, 1990). The time required to present chemosensory stimuli, however, usually precludes using more than two alternatives.

Test–retest reliability is not achieved at the expense of precision. Our measured thresholds show an approximately normal distribution, with a standard deviation of about 0.4 log units. The maximum-likelihood method is able to estimate thresholds that lie on a continuum, even though a small number of discrete stimulus concentrations are available for testing. Although a computer is re-

quired to run ML-PEST software, the widespread availability of inexpensive desktop or laptop computers makes this requirement much less of a barrier than in the past.

REFERENCES

Anliker, J. A., Bartoshuk, L., Ferris, A. N., & Hooks, L. D. (1991). Children's food preferences and genetic sensitivity to the bitter tastes of 6-*n*-propythiouracil (PROP). *American Journal of Clinical Nutrition*, **54**, 316-320.

Apter, A. J., Gent, J. F., & Frank, M. E. (1999). Fluctuating olfactory sensitivity and distorted odor perception in allergic rhinitis. *Archives of Otolaryngology: Head & Neck Surgery*, **125**, 1005-1010.

Bar-Hillel, M., & Wagenaar, W. A. (1993). The perception of randomness. In C. L. Gideon Keren (Ed.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 369-393). Hillsdale, NJ: Erlbaum.

Berkson, J. (1951). Why I prefer logits to probits. *Biometrics*, **7**, 327-339.

Berkson, J. (1953). A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function. *Journal of the American Statistical Association*, **48**, 565-600.

Berkson, J. (1955). Maximum-likelihood and minimum chi-square estimates of the logistic function. *Journal of the American Statistical Association*, **50**, 130-162.

Cain, W. S., Gent, J. F., Catalanotto, F. A., & Goodspeed, R. B. (1983). Clinical evaluation of olfaction. *American Journal of Otolaryngology*, **4**, 252-256.

Cometto-Muñiz, J. E., & Cain, W. S. (1995). Relative sensitivity of the ocular trigeminal, nasal trigeminal and olfactory systems to airborne chemicals. *Chemical Senses*, **20**, 191-198.

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

Cowart, B. J. (1989). Relationships between taste and smell across the adult life span. In C. Murphy, W. S. Cain, & D. M. Hegsted (Eds.), *Nutrition and the chemical senses in aging: Recent advances and current research needs* (Annals of the New York Academy of Sciences, Vol. 561, pp. 39-55). New York: New York Academy of Sciences.

Cowart, B. J., Yokomukai, Y., & Beauchamp, G. K. (1994). Bitter taste in aging: Compound-specific decline in sensitivity. *Physiology & Behavior*, **56**, 1237-1241.

Deems, D. A., Doty, R. L., Settle, R. G., Moore-Gillon, V., Shaman, P., Mester, A. F., Kimmelman, C. P., Brightman, V. J., & Snow, J. B., Jr. (1991). Smell and taste disorders, a study of 750 patients from the University of Pennsylvania Smell and Taste Center. *Archives of Otolaryngology: Head & Neck Surgery*, **117**, 519-528.

Doty, R. L., McKeown, D. A., Lee, W. W., & Shaman, P. (1995). A study of the test–retest reliability of ten olfactory tests. *Chemical Senses*, **20**, 645-656.

Doty, R. L., Snyder, P. J., Huggins, G. R., & Lowry, L. D. (1981). Endocrine, cardiovascular, and psychological correlates of olfactory sensitivity changes during the human menstrual cycle. *Journal of Comparative & Physiological Psychology*, **95**, 45-60.

Drewnowski, A., Henderson, S. A., & Shore, A. B. (1997). Genetic sensitivity to 6-*n*-propythiouracil (PROP) and hedonic responses to bitter and sweet tastes. *Chemical Senses*, **22**, 27-37.

Duffy, V. B., Cain, W. S., & Ferris, A. M. (1999). Measurement of sensitivity to olfactory flavor: Application in a study of aging and dentures. *Chemical Senses*, **24**, 671-677.

Fechner, G. T. (1860). *Elemente der Psychophysik*. Leipzig: Breitkopf und Härtel.

Gagnon, P., Mergler, D., & Lapare, S. (1994). Olfactory adaptation, threshold shift and recovery at low levels of exposure to methyl isobutyl ketone (MIBK). *Neurotoxicology*, **15**, 637-642.

Garcia-Perez, M. A. (1998). Forced-choice staircases with fixed step sizes: Asymptotic and small-sample properties. *Vision Research*, **38**, 1861-1881.

Green, D. M., & Swets, J. A. (1974). *Signal detection theory and psychophysics*. Huntington, NY: Robert E. Krieger. (Original work published 1966)

Grushka, M., Sessle, B. J., & Howley, T. P. (1986). Psychophysical evidence of taste dysfunction in burning mouth syndrome. *Chemical Senses*, **11**, 485-498.

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

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

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

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

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

Harvey, L. O., Jr. (1992). The critical operating characteristic and the evaluation of expert judgment. *Organizational Behavior & Human Decision Processes*, **53**, 229-251.

Harvey, L. O., Jr. (1997). Efficient estimation of sensory thresholds with ML-PEST. *Spatial Vision*, **11**, 121-128.

Hays, W. L. (1963). *Statistics for psychologists* (1st ed.). New York: Holt, Rinehart & Winston.

Krantz, D. H. (1969). Threshold theories of signal detection. *Psychological Review*, **76**, 308-324.

Lehrner, J. P., Brücke, T., Dal-Bianco, P., Gatterer, G., & Kryspin-Exner, I. (1997). Olfactory functions in Parkinson's disease and Alzheimer's disease. *Chemical Senses*, **22**, 105-110.

Lehrner, J. P., Kryspin-Exner, I., & Vetter, N. (1995). Higher olfactory threshold and decreased odor identification ability in HIV-infected persons. *Chemical Senses*, **20**, 325-328.

Linn, R. L. (Ed.) (1989). *Educational measurement* (3rd ed.). New York: Macmillan.

Linschoten, M. R. I., & Kroeze, J. H. A. (1991). Spatial summation in taste: NaCl thresholds and stimulated area on the anterior human tongue. *Chemical Senses*, **16**, 219-224.

Linschoten, M. R. I., & Kroeze, J. H. A. (1992). Bilateral taste stimulation: Spatial summation with weak NaCl stimuli. *Chemical Senses*, **17**, 53-59.

Linschoten, M. R. I., & Kroeze, J. H. A. (1994). Ipsi- and bilateral interactions in taste. *Perception & Psychophysics*, **55**, 387-393.

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

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

Mattes, R. D. (1988). Reliability of psychophysical measures of gustatory function. *Perception & Psychophysics*, **43**, 107-114.

Moll, B., Klimek, L., Eggers, G., & Mann, W. (1998). Comparison of olfactory function in patients with seasonal and perennial allergic rhinitis. *Allergy*, **53**, 297-301.

Murphy, C., & Cain, W. S. (1986). Odor identification: The blind are better. *Physiology & Behavior*, **37**, 177-180.

Murphy, C., Nordin, S., de Wijk, R. A., Cain, W. S., & Polich, J. (1994). Olfactory-evoked potentials: Assessment of young and elderly, and comparison to psychophysical threshold. *Chemical Senses*, **19**, 47-56.

Nordin, S., Murphy, C., Davidson, T. M., Quinonez, C., Jalowayski, A. A., & Ellison, D. W. (1996). Prevalence and assessment of qualitative olfactory dysfunction in different age groups. *Laryngoscope*, **106**, 739-744.

O'Mahony, M., Gardner, L., Long, D., Heintz, C., Thompson, B., & Davies, M. (1979). Salt taste detection: An R-index approach to signal-detection measurements. *Perception*, **8**, 497-506.

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

Pierce, J. D., Jr., Doty, R. L., & Amoore, J. E. (1996). Analysis of position of trial sequence and type of diluent on the detection threshold for phenyl ethyl alcohol using a single staircase method. *Perceptual & Motor Skills*, **82**, 451-458.

Rosenblatt, M. R., Olmstead, R. E., Iwamoto-Schaapp, P. N., & Jarvik, M. E. (1998). Olfactory thresholds for nicotine and menthol in smokers (abstinent and nonabstinent) and nonsmokers. *Physiology & Behavior*, **65**, 575-579.

SAS Institute (1998a). *StatView: StatView reference*. Cary, NC: SAS Institute.
SAS Institute (1998b). *StatView: Using StatView*. Cary, NC: SAS Institute.
Schlauch, R. S., & Rose, R. M. (1990). Two-, three-, and four-interval forced-choice staircase procedures: Estimator bias and efficiency. *Journal of the Acoustical Society of America*, **88**, 732-740.
Simpson, A. J., & Fitter, M. J. (1973). What is the best index of detectability? *Psychological Bulletin*, **80**, 481-488.
Stevens, J. C. (1995). Detection of heteroquality taste mixtures. *Perception & Psychophysics*, **57**, 18-26.
Swets, J. A. (1961). Is there a sensory threshold? *Science*, **134**, 168-177.
Swets, J. A. (1986a). Form of empirical ROC's in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin*, **99**, 181-198.
Swets, J. A. (1986b). Indices of discrimination or diagnostic accuracy: Their ROC's and implied models. *Psychological Bulletin*, **99**, 100-117.
Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Mahwah, NJ: Erlbaum.
Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review*, **68**, 301-340.
Taylor, M. M., & Creelman, C. D. (1967). PEST: Efficient estimates on probability functions. *Journal of the Acoustical Society of America*, **41**, 782-787.
Treutwein, B., & Strasburger, H. (1999). Fitting the psychometric function. *Perception & Psychophysics*, **61**, 87-106.
Urban, F. M. (1908). *The application of statistical methods to problems of psychophysics*. Philadelphia: Psychological Clinic Press.
Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, **33**, 113-120.
Weiffenbach, J. M., Baum, B. J., & Burghauser, R. (1982). Taste thresholds: Quality specific variation with human aging. *Journal of Gerontology*, **37**, 372-377.
Weiffenbach, J. M., Schwartz, L. K., Atkinson, J. C., & Fox, P. C. (1995). Taste performance in Sjögren's syndrome. *Physiology & Behavior*, **57**, 89-96.
Wetherill, G. B., & Levitt, H. (1965). Sequential estimation of points on a psychometric function. *British Journal of Mathematical & Statistical Psychology*, **18**, 1-10.

(Manuscript received November 11, 1999;
revision accepted for publication March 4, 2001.)