Chem. Senses 28: 817–826, 2003 DOI: 10.1093/chemse/bjg075

Human Odor Detectability: New Methodology Used to Determine Threshold and Variation

James C. Walker[1], Sandra B. Hall[2], Dianne B. Walker[1], Martin S. Kendal-Reed[1], Alison F. Hood[1] and Xu-Feng Niu[2]

[1]Sensory Research Institute, Florida State University, 1800 E. Paul Dirac Drive, Tallahassee, FL 32306-2741 USA and [2]Department of Statistics, Florida State University, Tallahassee, FL 32306-4330, USA

Correspondence to be sent to: James Walker, Sensory Research Institute, Florida State University, 1800 E. Paul Dirac Drive, B-340 NHMFL, Tallahassee, FL 32306-2741, USA. e-mail: jwalker@psy.fsu.edu

Abstract

Current ambiguity concerning the related issues of optimal means for measurement of odor sensitivity and the functional properties of the olfactory system hinders progress in basic and applied research on the human sense of smell. To address these needs, we selected *n*-amyl acetate (nAA) as a test odorant and developed a methodology in which participants (Ps) receive multiple presentations each session of several concentrations. Yes–no responses as to whether odor was detected are analyzed using binomial statistics, with the probability that a given proportion of yes responses (or greater) would occur by chance alone being treated as the inverse of detectability. Over the course of multiple sessions, this information is also used to maximize the collection of data in the peri-threshold region. Surprisingly, data collected over as many as 14 sessions were fit well by a single logistic regression model relating probability and concentration. Threshold concentrations, defined as those corresponding to a probability of 0.05, varied from 7.11 to 167.53 p.p.b. (v/v) for 11 Ps. Our approach and findings, if shown to be representative of other combinations of Ps and odorants, could accelerate the pace of research in human olfaction by providing a comprehensive operational definition of the limit of the olfactory system to detect odorant molecules.

Key words: acetate, amyl, human, olfactory, psychophysics, sensitivity

Introduction

Any attempt to understand the processing of odor information by the olfactory system should take into account the simplest measure of the performance limits of this system: odor threshold concentration. For example, any attempt to elucidate quality discrimination, or the intensity or qualitative aspects of odorant mixtures, must be based on experimental designs that take into account the odor potency of each chemical for each experimental participant. Though work in this area has been ongoing for over a century, it may be fairly stated that there is as yet no clear consensus as to how to quantify odor detectability. Historical summaries of work in olfactometry and psychophysical methods, respectively, may be found in contributions by Doty and Kobal (1995) and Prah *et al.* (1995); the former also contains a useful discussion of the relative merits of contemporary methods.

A useful though indirect measure of the lack of coherence on this question of threshold is the extremely large variation in values reported by different laboratories. Compilations such as that by Amoore and Hautala (1983), for example, include compounds for which inter-laboratory variation covers a range of ~10^6. Over roughly the past three decades, various critiques of odor psychophysical methodology have appeared (e.g. Hyman, 1977; Punter, 1983; Passe and Walker, 1985; Walker and Jennings, 1991). These have collectively provided a convincing case that much of the variation among laboratories is attributable to issues of experimental precision, especially concerning control of odorant concentration.

While there has been some recognition of the problem of inter-laboratory variation, there has been little investigation of inter-individual variation and the degree to which estimates of this source have been inflated by fluctuations in sensitivity over time for a given individual. A notable exception is the work of Stevens *et al.* (1988), who reported that the latter could range up to 10 000-fold and thereby erroneously inflate estimates of individual differences in sensitivity. The impression left by this important report, combined with a recent emphasis on cognitive influences on psychophysical responses (e.g. Distel and Hudson, 2001), seems to have

engendered a near abandonment of the idea that odor threshold is a quantitative and useful descriptive property of the olfactory system that can be measured in a scientifically valid way.

We suggest that this view should be re-evaluated and consideration given to the possibility that the prescription for the somewhat muddled situation outlined above might be an optimized methodology that attempts to address all of the shortcomings identified thus far. Ideally such a method should incorporate careful generation and presentation of odor stimuli, sufficient sampling of responses for each stimulus condition and well-reasoned steps for data processing. The latter component should provide not merely a threshold concentration value but a set of explicit tools for expressing variation, particularly that over time for a given individual. Additionally such a method should be readily communicated to different groups and amenable to comparisons between the human olfactory system and other chemical detector systems. In the present work, methodology substantially fulfilling these objectives was developed and then used to investigate some fundamental aspects of the human olfactory system.

Materials and methods

Participants

One group of seven individuals and a second group of five individuals served as participants (Ps). Gender and age information for the 11 Ps from whom we were able to obtain threshold data, as well as the twelfth P, are included in Table 2. All individuals were recruited through newspaper advertisement or flyers posted at various locations throughout Tallahassee, FL. At the initial interview, prospective Ps were questioned extensively as to any medical or exposure history associated with lowered olfactory sensitivity. Each also completed a health questionnaire designed to exclude individuals for whom any of the following were true: known abnormalities of smell or taste sensation; moderate to severe asthma or other serious chest disease; known allergy or sensitivity to *n*-amyl acetate (nAA) or other substances; pregnancy or suspected pregnancy; currently breast-feeding; cleft palate defects or surgery; broken nose; concussion; exposure to moderate or high concentrations of pesticides; nasal administration of non-therapeutic drugs (e.g. cocaine); severe heart disease; untreated high blood pressure; epilepsy; known HIV-positive status or AIDS sufferer; chronic adenoid or sinus disease; persistent nosebleeds; nasal polyps; conjunctivitis or extreme corneal sensitivity; untreated eye disease; cardiac pacemaker; any form of cancer; claustrophobia; epilepsy; smoker; or poor peripheral blood circulation. All participants were told that the purpose of the study was to examine sensory responses to odorous stimuli. Participants gave informed written consent and received financial compensation of $37.50 per test session for participating in the study, which was approved by the Institutional Review Board of the Florida State University.

Apparatus for odorant generation and presentation, and response measurement

An automated air dilution olfactometer (Walker *et al.*, 1990a,b) was used to generate vapor phase concentrations of nAA. This odorant is of value because it poses no known health risk at the concentration ranges used, is a poor nasal trigeminal stimulus (Walker *et al.*, 1990c; Warren *et al.*, 1994) and has been tested extensively in human and animal studies of olfactory function. This compound is also simple to use as an odorant, in part because oxidation to other compounds is a minimal concern. Reactivity of alkyl methyl esters toward oxygen at room temperature is negligible; with the reduced temperature we used, there is an even greater margin of safety. Clean air, at various volume flow rates, was passed through a glass saturator tube containing nAA and held at –5°C. This saturator temperature was chosen based on pilot testing indicating that this degree of cooling lowered the vapor pressure of nAA sufficiently that detectability was eliminated at the lower end of the fractions of vapor saturation that our system can produce. Since air is passed over the liquid surface, and no sparging occurs, no aerosols are generated. Air exiting the saturator was combined with a stream of clean air held at 25°C and 50% RH; the final volume flow rate of clean or odorized air delivered to the P was 43 l/min. Calibration of olfactometer output was achieved by use of an IR spectrometer (Miran 205B; Thermo Environmental Instruments; Franklin, MA). Real-time measurements of actual concentrations were made for the range from 0.05 (the lowest concentration quantitated by this instrument) to 1.66 p.p.m. Higher concentrations were not examined because of a concern that the saturator flow rates required for this concentration would result in incomplete saturation of air with nAA vapor. In addition, concentrations as high as 1.66 p.p.m. were not used to derive the concentration–probability function for even the least sensitive P. For the most sensitive P, concentrations approaching 1 p.p.b. were presented. The close relationship ($R^2 = 0.97$) between log vapor saturation and log p.p.m. was used to determine, by extrapolation, the concentrations corresponding to various fractions of vapor saturation below that yielding 0.05 p.p.m. This was done by measuring saturator flow rates at various programmed fractions of vapor saturation and using the regression equation relating vapor saturation to actual concentration to derive the latter.

Ps were tested in a well-lit, quiet room that was 3.8 m^2 in floor area and had a ceiling height of 2.6 m. Air within this room was 21 – 23°C and 20–45% relative humidity, and a true exhaust provided 10–15 air changes per h. When instructed by an auditory signal from the computer, the P pressed his/her face into a mask, the rim of which was inflated to produce a comfortable but snug fit around nose

and mouth. While the P was at the mask, recorded sounds of olfactometer operation mixed with 'white' noise were sent to the P through headphones. Though we have observed no evidence of auditory cueing of the P by instrument sounds, this masking stimulus is used as a precaution. We did not test for the possibility that this practice raised or lowered thresholds, but we suggest that this is unlikely given that the loudness was adjusted so as not to be aversive or distracting. Use of the inflated mask rim ensured that all air breathed by the P was from the olfactometer.

Breathing was measured by a pneumotachograph downstream of the P that quantified additions and subtractions (exhalations and inhalations) to the constantly supplied flow of 43 l/min. Instantaneous flow rate readings, sent to the computer at 100/s, were recorded for 8 s after the first inhalation onset. With the onset of the next exhalation, a flow valve near the facemask switched (over a period of ~0.2 s) from clean air to either odorized or (for control trials) clean air. That is, the transition from one stimulus condition to another is made while the P was breathing out. The dead volume of the mask varied among Ps but was sufficiently small (~100 ml) that replacement of this volume by olfactometer air required only ~0.1 s. This fact, the olfactometer flow rate of 43 l/min to the mask, and our tight control over degree of dilution of saturated vapor collectively ensured that the same odorant concentration was presented to the nasal cavity each time a given intensity was selected. This held across trials, sessions and Ps. Respiratory monitoring continues for an additional 8 s after onset of the subsequent inhalation. The masking stimulus is terminated and the P is then signalled by a recorded message to move from the mask and enter into a computer one of two responses: 'yes, odor was detected' or 'no, odor was not detected'. The P then waited ~90 s for the signal to begin another trial. During this inter-trial interval, the P is free to read and is able to relax until given an auditory cue by the computer that another trial is about to commence.

Procedure

Experimental design

Each of the original seven Ps was tested for at least twelve sessions, with a minimum of 48 h between sessions. The purpose of this approach was to ensure that we had sufficient sampling to assess intra-individual variation for each P. Analyses conducted on this group provided a clear basis for estimating the minimum amount of sampling (number of sessions) needed to arrive at a valid measure of odor threshold. These analyses also allowed us to develop a set of rules for defining, for each individual as testing progressed, when sufficient sessions had been completed. These rules were applied successfully, with minor modifications, to a second group of five Ps.

Apart from differences in amount of testing, the two groups of Ps were treated identically. For the first session, each P was presented with 15 clean air trials and 15 trials at each of four concentrations of nAA. These were specified in practice in terms of vapor saturation ($10^{-4.4}$, $10^{-4.1}$, $10^{-3.8}$, $10^{-3.5}$). Subsequent calibration determined that these values corresponded, respectively, to 0.044, 0.093, 0.20 and 0.43 p.p.m. (v/v). Presentation order of the five different stimulus conditions (four concentrations and clean air) was randomised, with the stipulation that no condition was presented more than three consecutive times. Each 75-trial session lasted ~2.3 h. Near the midpoint of each session, the P was offered a break of several min. We suggest that participation in such a study is not an onerous task for the P; we estimate that a total of 10 min or less was devoted to rendering simple yes–no judgements as to whether odor was present.

For the second and all other sessions, the range of concentrations presented was determined by responding to the lowest concentration on the previous test. For example, if the fraction of 'yes' responses was 3/15 for clean air trials and 6/15 for the lowest odorant concentration for session 1, binomial statistics would show that the likelihood of the level of responding (or greater) observed with the latter stimulus would be expected by chance alone 0.06 of the time. The proportion of yeses on clean air trials provides a direct measure of response bias for each session. This bias is taken into account by our use of this proportion, in the binomial probability calculations, as the assumed underlying distribution from which responses to odor stimuli are drawn. In cases where there were no 'yes' responses on clean air, a fraction of 0.5/15 was used instead of zero. If, for a session in which clean air yielded no 'yes' responses, the lowest odorant concentration also yielded no 'yes' responses, the following steps were taken. The probabilities associated with 'yes' frequencies of 0/15 and 1/15, given an assumed sampling distribution of 0.5/15, were determined and averaged. Once the probability associated with the lowest concentration was determined, it was evaluated against the rules illustrated in Table 1 to determine what change to make in the concentration range for the next session. For example, a probability of 0.06 would result in a lowering of the concentration range, for the next session, by 0.2 log unit.

Table 1 Rules for adjusting concentration range for one session based on responding to lowest concentration on prior session

Binomial probability for lowest concentration	How to change concentration range for next session
>0.5	raise by 0.1 log unit
≤0.5 and >0.1	lower by 0.1 log unit
≤0.1 and >0.01	lower by 0.2 log unit
≤0.01 and >0.001	lower by 0.3
≤0.001 and >0.0001	lower by 0.4
≤0.0001 and >0.000001	lower by 0.5

The interval between concentrations was held constant at 0.3 log unit. This approach provided a simple and generally successful means of ensuring that, once the P's dynamic range was approximated, data collection was concentrated on the transition from nearly perfect to nearly absent detection.

Data analyses

With each of the original seven Ps, we used logistic regression modeling to fit a function relating log concentration to the 4th root of probability. We chose logistic regression for several reasons. First, this approach is appropriate when, as is the case with our work, the probabilities being modeled derive from binomial responses. Secondly, this type of regression model makes the assumption that there is a monotonic relationship between X, the independent variable, and the probability p. Specifically we made the assumption, based on much prior work by others, that p would remain high until a sufficiently high concentration was reached, drop over ~1 log unit of concentration to very low values and then decline very slowly with further increases in concentration. Finally, our use of the generalized linear model (GLM) approach (see McCullagh and Nelder, 1989) avoided our making the assumption, necessary with traditional linear regression, that probabilities would be normally distributed for a given P-by-concentration combination.

The GLM approach we adopted enlarges the class of least square models in at least two ways. First, the distribution of responses (Y_i) for fixed predictors X (odorant concentration in this case) was assumed to be from the exponential family, which includes important distributions such as binomial, Poisson, exponential and gamma, in addition to the normal distribution. Secondly, it gives us the general form of

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta x$$

which is similar to traditional linear regression. This model can also be written in the form

$$p = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

which allows one to solve for probability with concentration (x) as the independent variable. Prior to applying the logistic regression approach, we transformed the probability to 4th root of probability. This step helped to reveal differences in detectability among supra-threshold intensities and aided visualization of the relationship. Statistical software (Splus 2000; Insightful Inc., Seattle, WA) was used to derive the parameters for the generalized linear model for each P. Outliers were found by plotting the fitted values; actual data points that were more than 2 SD away were defined to be outliers. The R^2 value was calculated using the formula:

$$R^2 = 1 - \frac{\sum (fitted)^2}{\sum (\sqrt[4]{prob} - mean(\sqrt[4]{prob}))^2}$$

The steps described above resulted in our being able to express odor detectability as the inverse of the (4th root of) likelihood that the observed level of responding, or better, would be observed due to chance alone. Defining threshold concentration was then a matter of simply selecting a criterion probability. We selected as our definition of threshold that concentration corresponding to a 4th root probability of 0.47 (untransformed probability of 0.05).

An R^2 value is perhaps the simplest means of assessing the degree to which data collected over a period of many weeks could be fit by a single model equation. The formula for this parameter is provided above. Two other approaches were examined to evaluate whether our approach of combining data over an extended period added, as one might expect from prior work, an unacceptable amount of intra-individual 'noise' to the concentration–probability function. First we calculated, for each concentration presented to a given P, the 95% confidence interval associated with the probability yielded by the logistic regression equation. To do this, we used a bootstrapping procedure (Efron, 1982). We first fit the model with the original data. Then, we obtained the fitted values and the residual values from this original model. In order to obtain a sample from this original data, we randomly selected, with replacement, from the residuals and added them to the original fitted values. These new data were fit and the fitted values were stored. This process was repeated 1000 times. The samples of fitted values were ordered from smallest to largest, separately for each concentration, and then samples 25 and 975 were chosen as the confidence interval limits.

Finally, we developed a procedure to estimate the minimum number of sessions of testing needed to obtain a valid threshold value for a given individual. For each subject, three sessions were randomly chosen. Then two models were fit for each individual subject. The first was

$$\sqrt[4]{prob} = \frac{\exp(\alpha_1 + \beta_1 * conc_1 + \alpha_2 + \beta_2 * conc_2)}{1 + \exp(\alpha_1 + \beta_1 * conc_1 + \alpha_2 + \beta_2 * conc_2)}$$

where α_1 refers to the intercept term related to the three randomly selected sessions only and α_2 refers to the intercept term related to the remaining sessions (after removing the three randomly chosen ones). Slope and concentration terms were similarly denoted as β_1 and β_2 and conc1 and conc2, respectively. The second model that was fit for each subject was

$$\sqrt[4]{prob} = \frac{\exp(\alpha + \beta * conc)}{1 + \exp(\alpha + \beta * conc)}$$

where α refers to the intercept term and β is the coefficient for the slope term. This is simply the model originally developed for all data for a given P. An *F* test was performed to determine if the original model was significantly different from that conducted with the sessions subdivided. If the *F* value showed the models were different then the process was repeated, with progressively more sessions being randomly selected, until the *F* value was statistically insignificant ($P > 0.05$). For each P, the number of sessions at which this first occurred was taken as the minimum number of 75-trial sessions needed for a valid determination of threshold.

Results

Original seven Ps

Our approach to the analysis of data appears to be quite useful for determining odor detectability. A single logistic regression model described well the relationship, for a given individual, between probability and concentration when data collected over a period of over three months were combined. Figures 1 and 2 depict the two Ps that represented the extremes in terms of the degree to which the concentration–probability data could be fit by a logistic regression equation; Ps 1 and 7 had R^2 values of 0.60 and 0.81, respectively. If one makes the assumption that 15 trials per concentration provides sufficient sampling for a given session, it appears from Figure 1 that the range of session-to-session variability was, for all individuals, sufficiently small to warrant the combining of data collected over 12–14 weeks of testing. Since a commonly observed, if seldom reported, observation from animal odor psychophysical studies is that sensitivity improves with prolonged testing, we tested for this trend. ANOVA determined that the P-by-session-by-concentration and session-by-concentration interactions were not significant. The absence of a trend toward increasing or decreasing sensitivity over time further supports the view that the concentration–probability data points around each logistic regression model are samples from a single, and surprisingly stable, sensitivity function.

If, as appears to be the case, the human olfactory system may be treated as a rather precise and stable instrument, some steps toward more detailed characterization may be taken. For example, one may examine uncertainty over the range of probabilities that signify the transition from nearly perfect to nearly absent detection. An effort in this direction is provided in Figure 3. For all seven of the original Ps, widths of the 95% confidence intervals are plotted as a function of 4th root of probability (inverse of detectability). Although there are differences among individuals, there is also a reasonably clear pattern for uncertainty to 'dip' at the approximate midpoint of the probability range. Only further work will reveal whether this pattern is robust and actually reflective of operating characteristics of the olfactory system. Ideally, this should involve new data being collected from additional individuals, in response to different stimuli, and processed using several approaches in addition to those which we propose. We selected as our definition of threshold that concentration corresponding to a 4th root probability of 0.47 (untransformed probability of 0.05); a vertical dashed line in Figure 3 denotes this probability value.

Admittedly the number of sessions we employed, and the number of weeks over which these sessions were conducted, appear somewhat extreme in view of the pattern of results. This approach reflects an attempt to ensure that sufficient sampling took place to allow complete characterization of variation over time for the most variable P. Since this appears to be the case, procedures described under 'Data analyses' were used to estimate the minimum number of sessions needed for valid determination of odor detect-

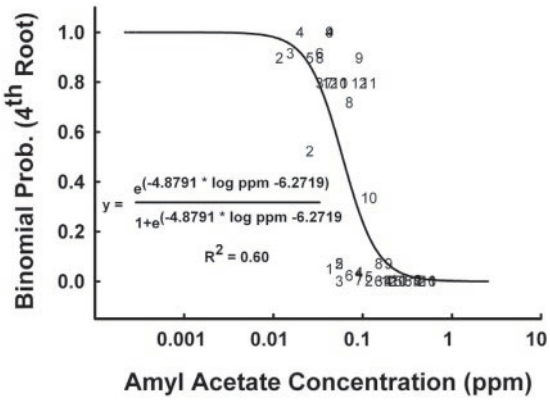

Figure 1 Scatter plot for one P (no. 1) showing the relationship between concentration of nAA and the 4th root of the binomial probability that the observed number of yeses, or greater, would be observed based on chance alone. Numerals denote session numbers. The logistic regression equation developed to model this relationship is shown and plotted. Of the original set of seven Ps, this individual exhibited the poorest fit between raw concentration–probability data points and the model.

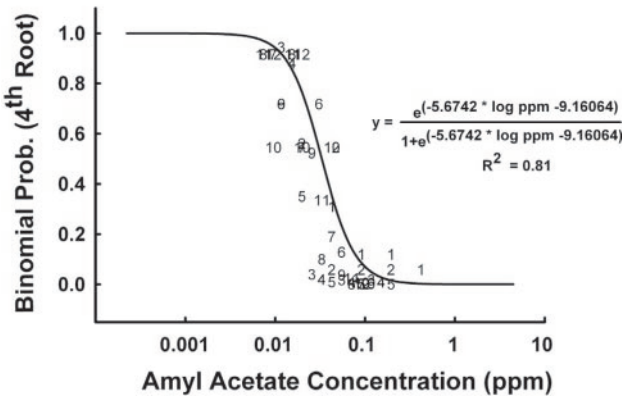

Figure 2 Same format as Figure 1. Of the original seven Ps, this individual (no. 7) exhibited the best fit between raw concentration–probability data points and the model.

ability. For one of the original seven Ps, five sessions were required; for the remaining Ps, only three sessions were needed. This information is summarized in Table 2.

Figures 4 and 5 depict the functions for Ps 3 and 6, the two individuals showing the lowest and highest thresholds, respectively. With our definition of threshold as that concentration corresponding to a probability of 0.05, the 4th root of which is 0.47, the thresholds for Ps 3 and 6 are 9.13 and 167.53 p.p.b. (v/v), respectively. Thus a range of less than 20-fold (< 1.3 log units) covered the sensitivity range for this initial group of seven individuals. Comparison of Figures 4 and 5 illustrates the value of examining uncertainty using a variety of approaches. Although the R^2 values for the two equations are similar, confidence intervals for that plotted in Figure 5 are much larger. This is attributable to the fact that this function exhibits a steeper slope. Since the probability declines (detectability increases) much more rapidly as concentration is increased, there is greater uncertainty associated with each concentration.

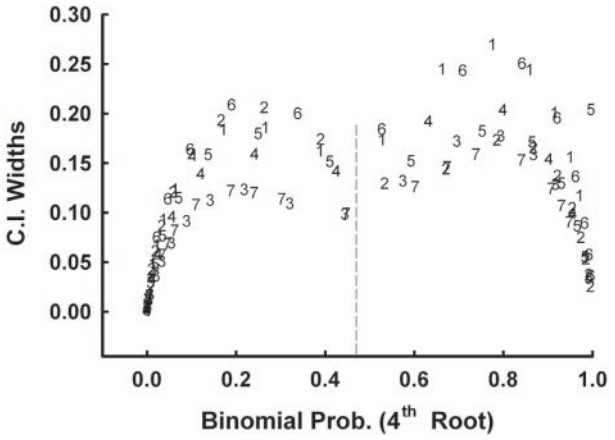

Figure 3 Plot of relationship between size of confidence interval and probability value. Data for the original seven Ps are shown, without regard for concentration, to explore possibility that uncertainty reaches a minimum at the approximate mid-point (on 4th root scale) between nearly perfect and nearly absent detection.

As discussed by Walker *et al.* (1999) and Kendal-Reed *et al.* (2001), conclusions about inter-individual differences are possible only after intra-individual variation has been accounted for. Since the foregoing has apparently achieved this objective for the original seven Ps, threshold concentration values may be calculated and compared. These are provided Table 2, along with R^2 values. Also included in this table is the concentration change, in log units, corresponding to a change in (4th root) probability from 0.05 to 0.80. Parameters such as this may prove useful in the process of quantifying, and then understanding, differences in the operating characteristics of individuals in terms of the processing of the signals received from the olfactory neuroepithelium. Our working hypothesis is that this slope measure is a characteristic of the olfactory system of each P. Thus, we would predict that, were an additional odorant to be tested with these same individuals, Ps 3 and 6 would represent extremes in terms of the rate at which detectability changes with concentration.

Table 2 Participant demographics and statistical descriptors of olfactory performance

P	Age at testing (years)	Gender	Threshold (p.p.b.)	R^2 for logistic regression model/no. of outliers removed	Minimum no.of sessions needed to define threshold	Slope measure (change in log concentration with 4th root probability change from 0.8 to 0.05)
1	26	M	62.77	0.6/0	3	0.85
2	52	M	13.82	0.71/1	5	0.83
3	23	F	9.13	0.74/0	3	0.91
4	19	F	15.09	0.8/0	3	0.57
5	35	F	18.90	0.78/3	3	0.65
6	48	F	167.53	0.71/0	3	0.61
7	20	M	25.51	0.81/0	3	0.77
8	20	F	15.88	0.95/2	ND	0.53
9	51	M	25.94	0.95/0	ND	0.36
10	44	M	36.49	0.9/0	ND	0.46
11	22	F	7.11	0.76/1	ND	1.55
12	42	F	could not be determined	no model constructed	NA	NA

ND, not determined; NA, not applicable.

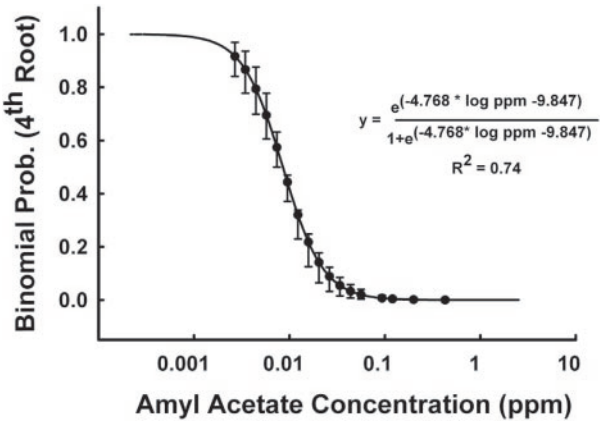

Figure 4 Plot of logistic regression model versus concentration for P (of original seven) exhibiting the greatest sensitivity (lowest odor threshold). Also shown are the confidence intervals associated with all of the concentrations presented. (Procedures for estimating confidence intervals are described under 'Data analyses'.)

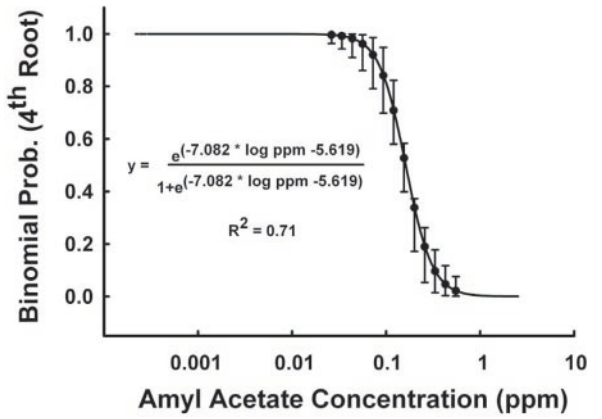

Figure 5 Same format as Figure 4. Plot of logistic regression model versus concentration for P (of original seven) exhibiting the least sensitivity (highest odor threshold).

Evaluation of procedure by testing of additional Ps

The procedures that we used with the seven original Ps, the basic findings in terms of modeling concentration–probability data and additional statistical analyses led us to conclude that our approach could be relied upon to provide valid data on odor detectability on new individuals with three to five test sessions. To partially evaluate this assumption, we recruited five additional Ps and tested each for nAA sensitivity using the procedures, for both data collection and analysis, outlined under 'Materials and methods'.

Based on our findings with the original seven Ps, we planned to test each of the five new Ps until at least four sessions were completed and the following 'stopping criteria' were satisfied: (i) at least four sessions were conducted, in each of which the 4th root binomial probability values ranged from ≤0.05 to ≥0.78; and (ii) logistic regression modeling, based on all sessions, yielded an R^2 value of ≥0.7. We were forced to modify the first criterion by the pattern of responding exhibited by one of the five new Ps. For this individual, a concentration range greater than that programmed for each session (0.9 log unit) was required for the transition from nearly perfect to nearly absent detection. Thus our first stopping criterion was modified so that the minimum of four sessions must include three sessions in each of which one concentration was included which yielded 4th root probabilities of ≤0.05, and three in each of which a concentration yielded probabilities of ≥0.78. A second of the five new Ps necessitated a third stopping criteria. This individual, even after 11 sessions, provided data that showed no evidence of a relationship between concentration and probability. These sessions occurred over a period of >20 weeks and spanned a concentration range of two log units. We are not able to offer an explanation for the pattern of results provided by this individual, though malingering should perhaps be considered among possible explanations. As a result of this presumably atypical individual, we have added the caveat that testing is not to be continued for an individual that provides data showing a non-monotonic concentration–probability relationship for ≥3 sessions.

The final set of criteria are thus as follows: (i) at least four sessions must be conducted; (ii) of these, three or more must include a concentration eliciting 4th root probabilities of ≤0.05, and three or more must include a concentration with probabilities of ≥0.78; (iii) logistic regression modeling of the concentration–probability function must yield $R^2 \geq 0.7$; (iv) testing must be terminated if three or more sessions exhibit a non-monotonic concentration–probability relationship. It should perhaps be emphasized that these are flexible criteria based on the data observed, as opposed to an assertion that a fixed number of test sessions will be sufficient to characterize any and all individuals in terms of odor detectability.

Application of the final stopping criteria resulted in, for four of the five new Ps, data that were quite similar to those obtained with the original seven Ps. These criteria prescribed from four to seven sessions. Since this exceeds the minimum number of sessions for all but one of the original seven Ps, we suggest that the final guidelines summarized above are somewhat conservative. Figures 6 and 7 show the Ps that exhibited, respectively, the lowest and highest rate of change in detectability with concentration. Whereas thresholds for the original seven Ps ranged from 9.13 to 167.53 p.p.b., those for these additional four Ps ranged from 7.11 to 36.49 p.p.b. (see Table 2). Logistic regression equations for all 11 Ps are plotted in Figure 8.

The clear advantages of collecting a large number of samples during a session, with each of several concentrations, must be weighed against at least two possible unwanted effects of prolonged testing. One concern is that trials might be spaced closely enough that the P loses sensitivity from one odorant trial to the next due to adaptation. To assess whether such an effect was produced by our method, we combined data for all but the first session for

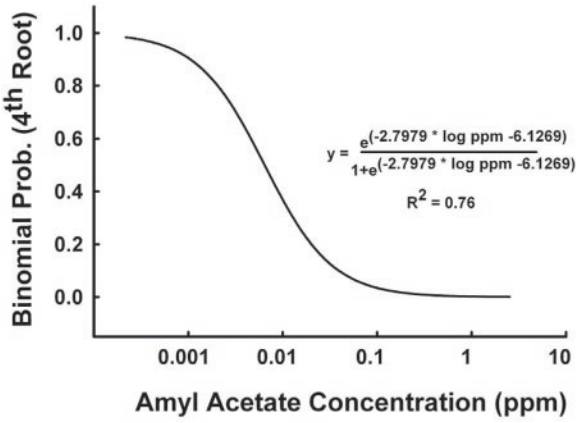

Figure 6 Plot of logistic regression model versus concentration for P (of second set of four) exhibiting the lowest rate of change in detectability with concentration.

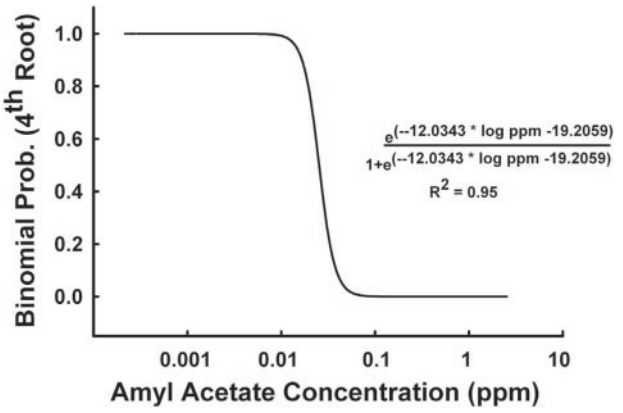

Figure 7 Same format as Figure 6. Plot of logistic regression model versus concentration for P (of second set of four) exhibiting the highest rate of change in detectability with concentration.

each P. All odorant trials (excluding the first trial of each session) were coded as to whether the preceding trial contained odorant or was a clean air control trial. Binomial testing was then used to evaluate the likelihood that the proportions of 'yes' responses seen on the two categories of odorant trials were different. The left half of Figure 9 depicts this comparison for each P. Also depicted in Figure 9 are the results of our effort to gauge the presence of a general fatigue over the course of the session. After excluding the first session for each P, odor trials within the first and last fifths of the 75-trial session were compared in terms of the proportion of 'yes' responses. As with the adaptation question, binomial testing was employed to evaluate differences. The patterns illustrated in both halves of Figure 9 provide little evidence that either adaptation or fatigue is a serious issue with the method we describe here.

Discussion

The methodology that we developed may offer some unique advantages due to the set of features that we have combined.

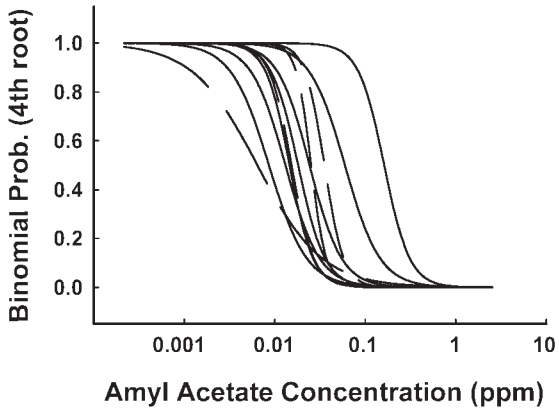

Figure 8 Summary of all 11 Ps in terms of regression models relating concentration to 4th root of probability. Plots for the original seven Ps are plotted in solid lines and dashed lines denote those for the four additional Ps.

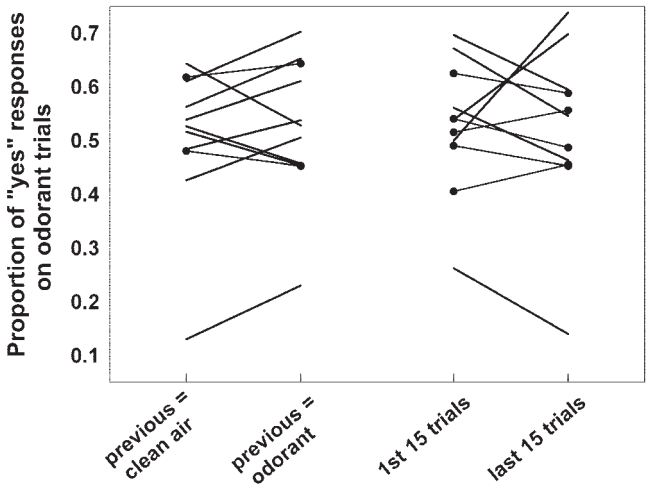

Figure 9 Summary of tests to assess possible adaptation (left panel) and fatigue. Data processing steps are described in 'Results'. Adaptation would be manifest as a lowered response when the previous trial contained odorant. Evidence for fatigue would be seen in a decline from early to late trials. Simple lines (without filled circles) denote statistically significant within-P changes.

Precisely controlled odorant concentrations are generated while holding constant other stimulus properties (e.g. total flow rate, RH). Odorant presentation is accomplished in a way that ensures that only olfactometer air is breathed and that odorant concentration is at full value with inhalation onset. Clear and simple instructions are given to Ps who, with apparently rare exception, produce samples of 'yes' and 'no' responses whose relative frequencies are systematically related to concentration. Each session, many samples (trials) are included per concentration; 15 appears to be a sufficient number. Repeated measurements from a given P were incorporated to ensure that intra-individual variation is not mistaken for that between individuals. Data are processed using steps that, though admittedly somewhat novel, are

transparent and provide quantitative measures of various aspects of each individual's concentration–probability function.

An unusual feature of our approach is that we tested each P, under essentially identical conditions, over a much greater number of days than is typical of human odor psychophysical studies. This emphasis derives from a long-term interest in the important question of the stability of odor function for a given individual. An oft-cited and very influential report (Stevens *et al.*, 1988) is to be lauded for early recognition of the importance of this question. Their findings suggested that intra-individual variation in odor sensitivity was on the order of 10 000-fold. This general finding, if replicated, would raise extremely troubling questions about any effort to integrate data collected over different days for a given individual, or from different individuals. Additionally, the fluctuations in function indicated by the Stevens *et al.* (1988) report would perhaps suggest little payoff from highly precise measurements of the olfactory system or from efforts to understand 'real world' responses to odors based on laboratory research.

Consistent with our prior work (Kendal-Reed *et al.*, 1998; Walker *et al.*, 1999), the present results indicate that intra-individual variation is far less than had been suggested by the seminal Stevens *et al.* (1988) work. As compared to our above-described prior work, the present study focused much more intensively on the peri-threshold region and incorporated repeated testing over a much longer period of time. This allowed us to define much more precisely the absolute threshold and to develop procedures for determining when an individual had been fully characterized in terms of odor detectability.

In evaluating differences among studies, parsimony leads us to have little enthusiasm for sorting out the degree to which various differences between the present and various prior procedures resulted in some quite different patterns of results. We think it is more important to develop a consensus among researchers as to how optimal procedures may be based on sound principles of stimulus control, sampling and statistical analysis of sources of variation. When results are in conflict, it would seem prudent to, at least initially, favor the more carefully collected data. As applied to the Stevens *et al.* (1988) report, this approach would lead to the hypothesis that the reported variation within individuals was primarily due to less than optimal attention to dimensions such as those listed above. Consistent with this notion, Stevens and Dadarwala (1993) note that apparent variation is greatly reduced by additional sampling. There are few data to which our finding of stable sensitivity to nAA may be compared. Wysocki *et al.* (1989) used this compound as a control odorant in studies of androstenone and found little or no change in sensitivity with repeated testing. The recent report (Dalton *et al.*, 2002) of improvements, of up to 100 billion-fold, in benzaldehyde sensitivity with repeated testing appears to incorporate no attempt to induce what is sometimes termed 'ultra-sensitivity' to nAA. The iso- form of this ester was employed only in tests conducted before and after repeated benzaldehyde detectability sessions.

If the assumption is made that the question of intra-individual variation has been dealt with satisfactorily with the present results, one can gain some idea of inter-individual variation in nAA sensitivity from Figure 8 and Table 2. A span of <25-fold separated the least and most sensitive Ps. The standardized value of 31 p.p.b. (Devos *et al.*, 1990) for nAA fell within the range of odor thresholds we report here: 7.11–167.53 p.p.b. Assessing the generality of our findings is difficult since prior work has not provided comparable measures of either intra- or inter-individual variation. Application of the method we describe and have validated, ideally in other laboratories, will determine the validity and generality of our findings as to slope differences among Ps and the degree of variation within and among individuals.

If the method we have developed survives the scrutiny of future work, the field of basic human olfaction would be able to advance on a number of important fronts. For example, one would now expect greater progress on the question of true mixture interactions. In the absence of rigorous testing of each combination of mixture component and P, effects that appear to be due to combinatorial actions arising from differences in chemical identities may actually reflect simply a failure to take into account the detectability of each odorant making up the test mixture. Similarly, accelerated progress may perhaps be expected in such categories of odor perception as odor quality discrimination, odor intensity discrimination and ability to identify a target odorant against a background of distractor chemicals.

Acknowledgements

Supported by the Center for Indoor Air Research, research funds from the Florida State University, a generous donation of instrumentation from the R.J. Reynolds Tobacco Co. and research funds from Philip Morris USA Inc. The technical assistance of D. Stan Warmath and Mark L. Thompson is gratefully acknowledged.

References

Amoore, J.E. and **Hautala, E.** (1983) *Odor as an aid to chemical safety: odor thresholds compared with threshold limit values and volatilities for 214 industrial chemicals in air and water dilution.* J. Appl. Toxicol., 3, 272–290.

Dalton, P., Doolittle, N. and **Breslin, P.A.S.** (2002) *Gender-specific induction of enhanced sensitivity to odors.* Nat. Neurosci., 5, 199–200.

Devos, M., Patte, F., Rouault, J., Laffort., P. and **Van Gemert, L.J.** (1990) Standardized Human Olfactory Thresholds. IRL Press, Oxford.

Distel, H. and **Hudson, R.** (2001) *Judgement of odor intensity is influenced by subjects' knowledge of the odor source.* Chem. Senses, 26, 247–251

Doty, R.L. and **Kobal, G.** (1995) *Current trends in the measurement of olfactory function*. In: Doty, R.L. (ed.), Handbook of Olfaction and Gustation. Marcel Dekker, New York, pp. 191–225.

Efron, B. (1982). *The bootstrap*. In Efron, B. (ed.), The Jackknife, the Bootstrap and Other Resampling Plans. Society for Industrial and Applied Mathematics, Philadelphia, PA, pp. 29–35.

Hyman, A.M. (1977) *Factors influencing the psychophysical function for odor intensity.* Sens. Process, 1, 273–291.

Kendal-Reed, M., Walker, J.C., Morgan, W.T. and **LaMachio, M.** (1998) *Human responses to propionic acid. I. Quantification of within- and between-subject variation in perception by normal and anosmic subjects.* Chem. Senses, 23, 71–82.

Kendal-Reed, M., Walker, J.C. and **Morgan, W.T.** (2001) *Investigating sources of response variability and neural mediation in human nasal irritation.* Ind. Air, 11, 185–191.

McCullagh, P. and **Nelder, J.A.** (1989) Generalized Linear Models. New York, Chapman & Hall.

Passe, D.H. and **Walker, J.C.** (1985) *Odor psychophysics in vertebrates*. Neurol. Biobehav. Rev., 9, 431–467.

Prah, J.D., Sears. S.B. and **Walker, J.C.** (1995) *Modern approaches to air-dilution olfactometry*. In: Doty, R.L. (ed.), Handbook of Olfaction and Gustation. Marcel Dekker, New York, pp. 227–255.

Punter, P.H. (1983) *Measurement of human olfactory thresholds for several groups of structurally related compounds.* Chem. Senses, 7, 215–235.

Stevens, J.C. and **Dadarwala, A.D.** (1993) *Variability of olfactory threshold and its role in the assessment of aging.* Percept. Psychophys., 54, 296–302.

Stevens, J.C., Cain, W.S. and **Burke, R.J.** (1988) *Variability of olfactory thresholds.* Chem. Senses, 13, 643–653.

Walker, J.C. and **Jennings, R.A.** (1991) *Comparison of odor perception in humans and animals.* In Laing, D.G., Doty, R.L. and Breipohl, W. (eds), The Human Sense of Smell. Springer-Verlag, New York, pp. 261–280.

Walker, J.C., Kurtz, D.B., Shore, F.M., Ogden, M.W. and **Reynolds, J.H.** (1990a) *Apparatus for the automated measurement of the responses of humans to odorants*. Chem. Senses, 15, 165–177.

Walker, J.C., D.B. Kurtz and **Shore, F.M.** (1990b) *Apparatus for assessing responses of humans to stimulants*. US patent 4,934,386, dated June 19, 1990.

Walker, J.C., Reynolds, J.H., Warren, D.W. and **Sidman, J.D.** (1990c) *Responses of normal and anosmic subjects to odorants.* In Green, B.G., Mason, J.R. and Kare, M.R. (eds), Chemical Senses. Vol. 2. Irritation. Marcel Dekker, New York, pp. 95–121.

Walker, J.C., Kendal-Reed. M. and **Morgan, W.T.** (1999) *Accounting for several related sources of variation in chemosensory psychophysics.* In Bell, G.A. and Watson, A. (eds), Tastes and Aromas: The Chemical Senses in Science and Industry. University of New South Wales Press, Sydney, pp. 105–113.

Warren, D. W., Walker, J.C., Drake, A.F. and **Lutz R.W.** (1994) *Effects of odorants and irritants on respiratory behavior.* Laryngoscope, 104, 623–626.

Wysocki, C.J., Dorries, K.M. and **Beauchamp, G.K.** (1989) *Ability to perceive androstenone can be acquired by ostensibly anosmic people.* Proc. Natl Acad. Sci. USA, 86, 7976–7978.

Accepted October 28, 2003