A Study of the Test-retest Reliability of Ten Olfactory Tests

Richard L. Doty, Donald A. McKeown, W. William Lee and Paul Shaman

Smell and Taste Center and Department of Otorhinolaryngology: Head and Neck Surgery, University of Pennsylvania Medical Center, 3400 Spruce Street and Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA

Correspondence to be sent to: R.L. Doty, Smell and Taste Center, 5 Ravolin Pavilion, University of Pennsylvania Medical Center, 3400 Spruce Street, Philadelphia, PA 19104, USA

Abstract

Ten tests of olfactory function (including tests of odor identification, detection, discrimination, memory, and suprathreshold odor intensity and pleasantness perception) were administered on two test occasions to 57 subjects ranging in age from 18 to 83 years. The stability of the average test scores was determined across the two test sessions for 14 measures derived from these 10 tests and for subcomponents of the Japanese T&T olfactometer threshold test. In addition, the test-retest reliability (Pearson *r*) of each test measure was established. With the exception of a response bias measure, the average test scores did not differ significantly across the two test sessions. Statistically, the reliability coefficients of the primary test measures fell into three general classes bound by the following *r* values: 0.43–0.53; 0.67–0.71; 0.76–0.90. Detection threshold values were more reliable than recognition threshold values; those based upon a single ascending presentation series were much less reliable than those based upon a staircase procedure. The relationship between test length and reliability was examined for several of the tests and mathematically modeled. For example, within the staircase series incorporating the odorant phenyl ethyl alcohol, reliability was related (R^2 = 0.984) to the number of reversals included in the threshold estimate by a function derived from the Spearman–Brown formula; namely, reliability = 0.455* # reversals/[1 + 0.455 (# reversals − 1)]. Reversal location, *per se*, had little influence on reliability. Overall, this study suggests that (i) considerable variation is present in the reliability of olfactory tests, (ii) reliability is a function of test length, and (iii) caution is warranted in comparing results from nominally different olfactory tests in applied settings since the findings may, in some instances, simply reflect the differential reliability of the tests.
Chem. Senses 20: 645–656, 1995.

Introduction

Given the importance of olfaction to humans, it is not surprising that numerous procedures have been developed, since the pioneering human olfactory studies of Valentin (1848), Toulouse and Vaschide (1899), Zwaardemaker (1889) and Proetz (1924), to test the sense of smell in clinical, academic and industrial settings. Included in such procedures are tests of olfactory sensitivity (e.g. odor detection and recognition thresholds), odor discrimination, odor identification, odor memory and suprathreshold scaling of odor intensity and pleasantness (for reviews, see Wenzel, 1948; Harper *et al.*, 1968; Köster, 1975; Takagi, 1989; Doty, 1991, 1992, 1995; Cain *et al.*, 1992; Doty and Kobal, 1995).

Despite the development of a wide range of such tests, little is known about their psychometric properties. For example, even basic information about the reliability of most olfactory tests is lacking and the handful of studies that have addressed this issue have done so only for a few types of tests (e.g. tests of odor identification and detection; see Punter, 1983; Doty *et al.*, 1984b, 1989; Cain and Gent, 1991). Psychometrically, reliability is a prerequisite to validity and, for this reason, tests which are unreliable have little utility. Thus, when nominally distinct tests of disparate reliability are administered, differences in test results could simply reflect differences in test reliability, rather than differences in underlying physiological or psychological processes purportedly measured by the tests (see Chapman and Chapman, 1978).

In this study, we examined the reliability and stability of 14 measures derived from 10 olfactory tests, including tests of odor detection, identification, discrimination, memory, and suprathreshold intensity and pleasantness perception. Additionally, the relationship between test length and test reliability was examined. Since shorter tests minimize fatigue, test time and cost, knowledge of the functional relationship between test length and reliability may be of practical use in establishing the optimal length of a given test.

Materials and methods

Subjects

Fifty-seven healthy subjects participated [21 men, 36 women; mean (SD) age = 42.65 (19.03); mean (SD) years of education = 15.60 (1.71); number of current, past and never smokers = 8, 28 and 21, respectively]. All scored well on the Picture Identification Test (PIT), a test designed to detect cognitive deficits which would interfere with non-olfactory components of olfactory tests such as the UPSIT [mean PIT (SD) = 39.91 (0.34)] (e.g. Doty *et al.*, 1987). The participants were students and staff of the University of Pennsylvania, and healthy, ambulatory participants recruited from the Philadelphia Center for Older People and the Medford Leas retirement community (Medford, NJ). Each subject received $25.00 for participation in this and a related study (Doty *et al.*, 1994) and, in accord with the University of Pennsylvania's Committee on Studies Involving Human Beings, provided informed written consent.

Test procedures

A battery of 10 tests of olfactory function was administered to each subject on two test occasions, separated from one another, on average, by about 2 weeks (median days between test occasions = 12, interquartile range = 18). These tests, which are described in detail below, were administered in random order on the first test occasion. For a given subject, the order of testing on the second test occasion was the same as on the first.

University of Pennsylvania Smell Identification Test (UPSIT)

This standardized test (commercially available as the Smell Identification Test™, Sensonics, Inc, Haddon Hts, N.J.) is the most widely used olfactory test in North America, having been administered to at least 35 000 persons in the last decade. In this test, a subject is required to identify, in a four-alternative multiple choice format, each of 40 odorants presented on micro-encapsulated 'scratch and sniff' labels. For example, one of the test items reads, 'This odor smells most like: (a) chocolate; (b) banana; (c) onion; or (d) fruit punch'. The subject must provide a response even if no odor is perceived (i.e. the test is forced-choice). The dependent measure is the number of items correctly answered. Specifics of this test are described in detail elsewhere (Doty *et al.*, 1984a, b, 1989).

Modular Smell Identification Test

This forced-choice test, also termed the Cross-Cultural Smell Identification Test™ (CC-SIT), is comprised of a subset of UPSIT items and is designed to be administered in less than 5 min (Doty *et al.*, 1995). Each subject is required to identify each of 12 microencapsulated odorants in a single-booklet, four-alternative, multiple-choice format. The number of items out of 12 that were answered correctly served as the dependent measure.

Single ascending series butanol odor detection threshold test

This test, which is described by Cain *et al.* (1988) and Stevens *et al.* (1988), consists of 12 ternary aqueous dilution steps of *n*-butanol (from a 4% initial dilution mixture) presented in ascending order in a two-alternative forced-choice paradigm. The lowest concentration at which a subject correctly indicated which of two plastic squeeze bottles, one containing the odorant and the other the diluent, produced the stronger odor on five consecutive trials served as the threshold measure (see Cain and Rabin, 1989).

Phenyl ethyl alcohol single staircase odor detection threshold test

This test determines a detection threshold value for the rose-smelling odorant phenyl ethyl alcohol by using a modified single staircase procedure described in detail elsewhere (Doty *et al.*, 1984b; Deems and Doty, 1987). In this study, the staircase was begun at the −6.00 log concentration step of a half-log step (vol/vol) dilution series extending from −10.00 log concentration to −2.00 log concentration. The odorant was increased in full log steps until correct detection occurred on five sets of consecutive trials at a given concentration. If an incorrect response was given on any trial, the staircase was moved upward a full log step. When a correct response was made on all five trials, the staircase was reversed and subsequently moved up or down in 0.50-log increments or decrements, depending upon the subject's performance on two pairs of trials at each concentration step. The geometric mean of the first four staircase reversal points following the third staircase reversal was used as the threshold measure. In the few occasions where a ceiling effect or basement effect occured, the threshold estimate was calculated as the highest or lowest concentration, respectively, detected.

Single series phenyl ethyl methyl ethyl carbinol odor detection threshold test

This test establishes a measure of detection threshold for the odorant phenyl ethyl methyl ethyl carbinol (PEMEC) by using squeeze bottles (see Amoore and Ollman, 1983). A $\log_2$ dilution series analogous to that used in the *n*-butanol threshold test above was presented in a single ascending method of limits series, with the exception that only three correct pairs of trials at a given concentration were required to define the threshold value.

Odor recognition memory test

In this 12-trial test, a microencapsulated 'target odorant' is presented to a subject on a given trial, followed by four odorants from which the subject is instructed to select the one identical to the target stimulus. On one-third of the trials, a 10-s interval was interspersed between the sampling of the target stimulus and the presentation of the first of the four alternatives. On another third, a 30-s interval was enforced, whereas on the other third of the trials, a 60-s period intervened. The number of trials in which the target odor was correctly identified served as the primary dependent measure.

Odor discrimination test

In this test, a subject was presented with 16 sets of three microencapsulated odorants (two same, one different) on separate pages of a cardboard test booklet (see Smith *et al.*, 1993). The subject is required to select the 'odd' or 'different' odor of the three. The odorants of a triad were preselected to be equivalent in average perceived intensity, as determined from nine-point category scale ratings (see Doty *et al.*, 1984b, for details). The number of triads in which the different stimulus is correctly reported served as the dependent measure.

Yes–No odor identification test

In this test, described by Corwin (1989), each of 20 odorants is presented twice—once with a descriptor that correctly describes the smell and once with a descriptor that does not. The subject's task is to report 'yes' or 'no' as to whether the odor smells like the given descriptor. In addition to the percent of trials judged correctly, two measures derived from signal detection theory serve as dependent variables: namely, *d′* and Cl (Snodgrass and Corwin, 1988). *d′* reflects the sensory sensitivity and Cl the response bias (i.e. the criterion an individual uses to make a decision as to whether or not a stimulus is present).

Suprathreshold amyl acetate odor intensity and odor pleasantness rating tests

In this test, 100-ml glass sniff bottles containing different concentrations of amyl acetate (−1.00, −2.00, −3.00 and −4.00 log vol/vol) diluted in USP grade light mineral oil are presented to the subject. Each of the four stimuli was presented five times apiece, in counterbalanced order. The subject was required to rate the perceived intensity and pleasantness of each stimulus on anchored nine-point category scales (for intensity, 1 = no smell, 9 = extremely strong; for pleasantness, 1 = dislike extremely, 9 = like extremely). Two measures are calculated for the odor intensity ratings: (a) the slope of the concentration/intensity function (following log transformation of the intensity ratings) obtained from a least-squares linear regression analysis; and (b) the overall mean of the intensity ratings. For the pleasantness assessment, only the mean of the pleasantness ratings was used as the dependent measure, since (i) no single function has been found which uniformly fits the response/concentration data for the majority of the subjects and (ii) pleasantness ratings are relatively flat over a wide range of amyl acetate concentrations (Doty, 1975).

T&T olfactometer

The T&T olfactometer test is routinely administered to thousands of patients each year by otorhinolaryngologists in Japan and is the only olfactory test for which Japanese physicians receive insurance reimbursement. This test consists of bottles containing five odorants, each diluted into eight log-step concentration series using either propylene glycol or Nujol oil (for details, see Takagi, 1989; Yoshida, 1984). The stimuli are iso-valeric acid, skatole, β-phenyl ethyl alcohol, Γ-undecalactone and methyl cyclopentenolone (cyclotene). The stimulus concentrations were presented in an ascending series in a non-forced choice situation and sniffed from strips of blotter paper dipped into the odorant solutions. The concentration at which a stimulus was first noticed (but not recognized) was defined as the detection threshold, whereas the concentration where a qualitative sensation was recognized was defined as the recognition threshold.

Mathematical fitting of test length/reliability functions

A test's reliability increases as its domain becomes more thoroughly sampled. Therefore, we examined the relationship between test length and reliability for those tests where such an evaluation was possible and determined the goodness-of-fit of several mathematical models. Three models were chosen for evaluation and were fit using the nonlinear estimation module of SYSTAT (Wilkinson, 1990).

The first model was based on the Spearman–Brown formula, an equation that is widely used for estimating the reliability of a test when its length is changed (see, e.g. Guilford, 1954; Magnusson, 1967). Suppose a test of length n has test-retest reliability, r. What would the reliability be if the test length is increased by m times? The Spearman–Brown formula predicts that the new test with length $m * n$ will have reliability

$$\frac{mr}{1 + (m - 1)\,r}$$

Since this formula is derived from the fundamental axioms underlying the calculation of the reliability coefficient [see, for example, Guilford's (1954) derivations], it also represents the simplest relationship possible between reliability and test length given the assumptions of the derivation. Specifically, we can re-write the formula such that m, the number of times test length is increased or decreased, is replaced by n, test length. Formally,

$$r = \frac{n\mathrm{r}_1}{1 + (n{-}1)\,\mathrm{r}_1},$$

where r is the dependent variable (reliability), n is the independent variable (test length), and r_1 is a free parameter[1] representing the reliability of the test when its length is 1. We chose this formula not only because of its grounding in psychological test theory, but because of its accurate prediction, in an earlier study of UPSIT fractions, of split-half and other intratest reliability coefficients (Doty *et al.*, 1989).

The second model we determined goodness-of-fit for was the logarithmic function,

$$r = \mathrm{b}\log(n) + \mathrm{c},$$

where b and c are free parameters. Since test reliability increases in a negatively accelerated manner (Cronbach, 1960), the logarithmic function was expected, *a priori*, to provide reasonable fits to some of the reliability/test length data. Note, however, that the logarithmic model has one more free parameter than the Spearman–Brown model. Thus, if the Spearman–Brown model accounts for (or fits) the data as well as the logarithmic model, it would be considered as the better (or more parsimonious) description of the relationship between reliability and test length.

The third model we evaluated was the linear function,

$$r = \mathrm{b}\,(n) + \mathrm{c},$$

where b is the slope of the function and c is the y-intercept. This straight-forward model might be expected to fit some reliability/test length data which exhibit monotonicity and minimal curvilinearity, and which do not show a tendency towards asymptote at longer test lengths (e.g. data from very short tests with few items). Since a linear model has the same number of free parameters (i.e. 2) as the logarithmic model, a direct comparison of the degree to which these two models fit a common set of data can be made.

We also modeled the data by adding a 'saturation' assumption to the aforementioned models, since, in some cases, the reliability/test length data appeared to flatten at a well-defined break point. While a third free parameter (i.e. saturation point) is needed for these models, such models allow for the specification of the saturation point. Thus, in

Table 1 Mean (SD) test scores for measures assessed in this study. See text for description of units of measurements. Only the Yes–No ID bias score differed significantly across test sessions [$t(56) = 4.91$, $P < 0.001$] Tests are ordered alphabetically

	Test 1	Test 2
Butanol detection threshold	6.72 (2.65)	5.96 (2.01)
Modular UPSIT	10.82 (1.76)	10.89 (2.13)
Odor discrimination	10.77 (2.79)	11.04 (2.84)
Odor memory	8.12 (2.85)	7.91 (2.89)
PEA detection threshold (log v/v)	–6.38 (1.63)	–6.28 (1.74)
PEMEC detection threshold	14 30 (9.66)	13.07 (11.87)
Suprathreshold intensity rating (mean)	4.40 (0.88)	4.39 (0.90)
Suprathreshold intensity rating (slope)	0.15 (0.06)	0.16 (0.06)
Suprathreshold pleasantness rating (mean)	4.63 (0.84)	4.56 (0.83)
T&T detection composite mean	–0.58 (1.24)	–0.47 (1.36)
T&T recognition composite mean	1.36 (0.97)	1.33 (1.14)
UPSIT	35.84 (5.49)	35.63 (5.83)
Yes–No ID (# correct)	32.74 (4.26)	32.04 (4.81)
Yes–No ID bias (CI)	–0.29 (0.74)	–0.78 (0.80)
Yes–No ID (d′)	3.22 (1.42)	3.26 (1.45)

future applications of a test in which such a model provides a strong fit, the tests could be shortened to the point where the saturation occurs without altering their degree of reliability.

Results

Stability of average test scores

The mean (± SD) test score values of the 14 primary test measures derived from the 10 olfactory tests are presented in Table 1 for the two test occasions; those of the T&T subtests are presented in Table 2. *t*-Tests applied to each test measure (Bonferroni corrected *P*-values) found a significant difference across the two test occasions only for the response bias measure derived from the yes-no test. In this case, the subjects adopted a more liberal criterion (i.e. evidenced increased numbers of false alarms) on the second test occasion than on the first.

Table 2 Mean (SD) threshold values for T&T olfactometer subtests. No significant differences were observed between session 1 and session 2 test scores

	Test 1	Test 2
T&T detection thresholds		
β-Phenyl ethyl alcohol	–0.63 (1.48)	0.56 (1.46)
Cyclotene	–0 54 (1.39)	–0.28 (1.59)
Γ-undecalactone	–0.60 (1.40)	–0.37 (1.69)
Isovaleric acid	–0.49 (1.27)	–0.53 (1.49)
Skatole	–0.63 (1.52)	–0.60 (1.66)
T&T recognition thresholds		
β-Phenyl ethyl alcohol	1.07 (1.87)	0.98 (1.78)
Cyclotene	1.81 (1.39)	1.68 (1.65)
Γ-undecalactone	1.95 (1.83)	2.44 (1.98)
Isovaleric acid	1.18 (1.39)	0.81 (1.51)
Skatole	0.81 (1.91)	0.75 (1.98)

Reliability of test measures

The test-retest reliability coefficients are presented in Tables 3 and 4 for the 14 primary test measures and the subcomponents of the T&T olfactometer, respectively. For the primary test measures, the *r* values ranged from 0.43 to 0.90, with a median value of 0.69. For the measures of the T&T subcomponents, these coefficients ranged from 0.33 to 0.69, with the detection threshold values evidencing higher reliability than the recognition threshold values. With the exception of the cyclotene and Γ-undecalatone recognition threshold values (Table 2; $P < 0.01$ and 0.05, respectively), all *r* values were significant beyond the 0.001 alpha level.

We tested the statistical significance among the reliability

Table 3 Test-retest reliability coefficients (Pearson *r*'s) for test measures evaluated in this study Measures ranked according to magnitude of *r* values. All *r* values significant beyond the $P < 0\ 001$ level $n = 57$. The values within a rectangle did not differ statistically from one another

UPSIT	0.90
PEA single staircase detection threshold	0.88
Suprathreshold pleasantness rating (mean)	0.78
Suprathreshold intensity rating (mean)	0.76
Modular UPSIT	0.71
T&T composite mean	0.71
PEMEC detection threshold	0.70
Yes-no ID (# correct)	0.69
Odor memory	0.68
Suprathreshold intensity rating (slope)	0 68
Yes-no discrimination (*d'*)	0.67
T&T composite mean	0.53
Yes-no bias (CI)	0.51
Butanol detection threshold	0.49
Odor discrimination (# correct)	0.43

Table 4 Test-retest reliability coefficients (Pearson *r*'s) for T&T olfactometer test measures evaluated in this study. Measures ranked according to magnitude of *r* values All *r* values significant at $P < 0\ 001$, with the exception of the recognition threshold for cyclotene ($P < 0.01$) and for Γ-undecalactone ($P < 0.05$)

T&T detection thresholds	
Skatole	0.71
Isovaleric acid	0.69
Γ-undecalactone	0.68
β-Phenyl ethyl alcohol	0.57
Cyclotene	0 56
T&T recognition thresholds	
Isovaleric acid	0 45
β-Phenyl ethyl alcohol	0.44
Skatole	0.42
Cyclotene	0.37
Γ-undecalactone	0.33

coefficients using a *z*-statistic developed specifically for this purpose[2].

The *r* values for the primary test measures grouped into three general classes, the members of which did not differ significantly from one another at the nominal $P < 0.05$ probability level. In the case of the T&T olfactometer subtest (Table 4), the recognition threshold reliability coefficients were consistently lower than the detection threshold reliability coefficients. However, despite this consistency, only the 0.71 reliability coefficient for the skatole detection threshold differed significantly ($P < 0.05$) from the other measures (namely, all of the recognition threshold measures).

Relationship of reliability coefficients to test-retest interval

No meaningful relationships were found, for any of the measures, between (a) the magnitude of differences between the two test scores and (b) the time, in days, between their administration (Pearson *r*'s; *P*'s > 0.20).

Relationship of reliability coefficients to test length

The reliability/test length data points and best-fit functions are shown in Figure 1. Test length reflects all of the test items or reversals up to the indicated test length. The sum of squared deviations (SSD) and percentage of variance accounted for (R^2) by each of the models tested in this study are presented in Table 5. Since correlation coefficients based upon a small number of test scores are unstable, we fit the models to reliability coefficients data based upon a minimum of five consecutive test items. In the case of the phenyl ethyl alcohol detection threshold test, all reversals were used in the model fitting since the first reversal was based upon a number of stimulus presentations[3].

It is apparent from Figure 1 that, in most cases, strong relationships existed between test length and reliability. It is also apparent from Table 5 that no one model uniformly provided the best fit to the various data sets. For example, for three tests (UPSIT, mean pleasantness rating, and yes-no ID bias), the non-saturation logarithmic model accounted for most of the variance in the data. In two cases, the Spearman–Brown non-saturation model seems to best characterize the data (odor discrimination and PEA detection threshold tests), whereas in two cases (odor memory and intensity rating), a simple linear function best described the data. In the case of the MODSIT and the yes/no odor ID tests, the linear saturation model seemed to provide the best fit[4].

To address the issue as to whether reliability was related to reversal number or position within the PEA staircase series, we computed reliability coefficients for (i) the first staircase reversal (analogous to the data point used for determining the threshold value in the butanol and PEMEC threshold tests), (ii) each of the other staircase reversals, and (iii) successive combinations of 2, 3, 4, 5, 6 and 7 staircase reversal points. Data from subjects evidencing

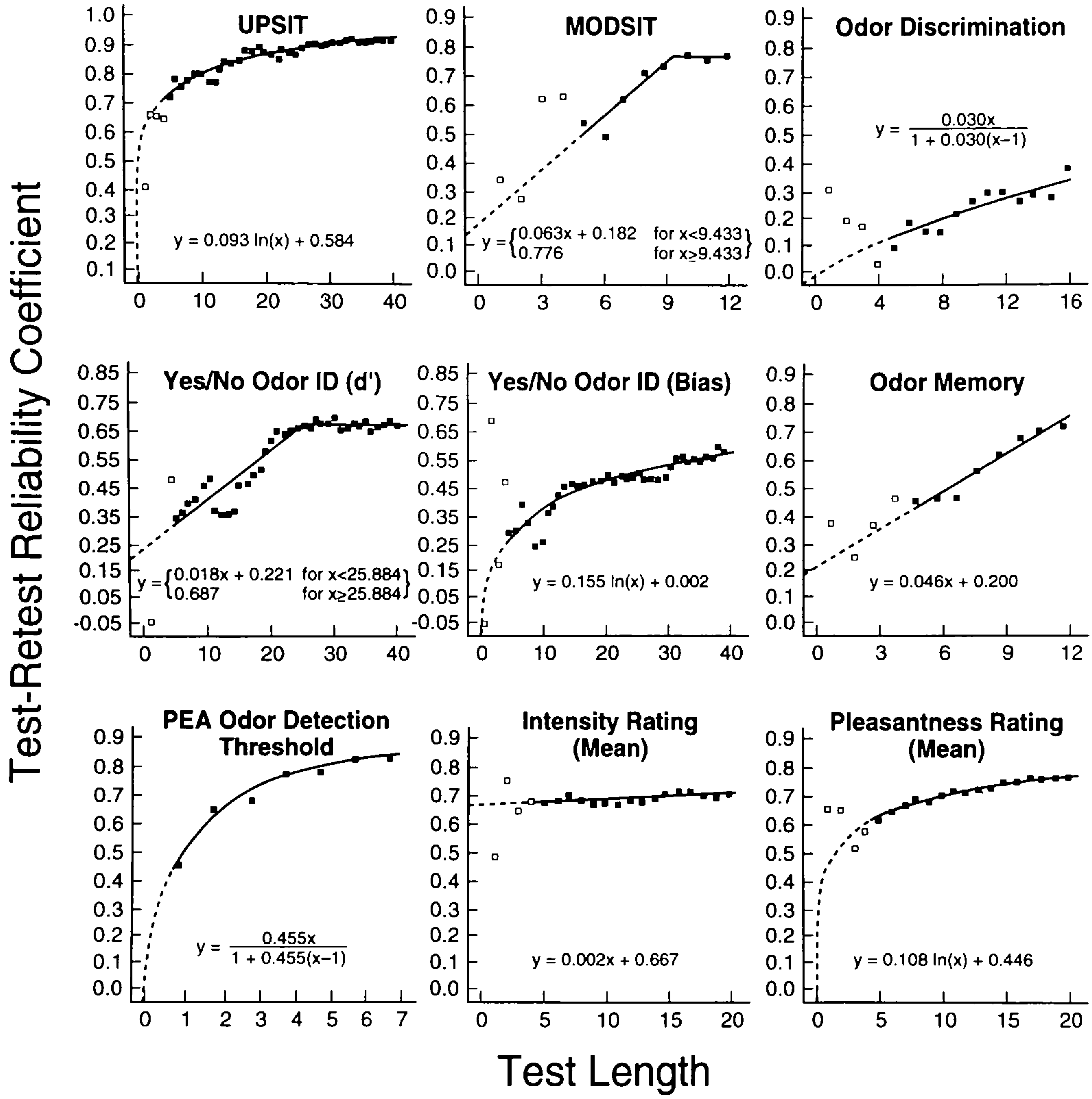

Figure 1 Relationship of reliability to cumulative test length for test measures amenable to such an evaluation. See text for details and Table 5 for R^2 values of fitted functions. Functions were modeled only on the filled squares.

ceiling or basement effects on either the first or second test occasion were omitted from these analyses. As can be seen in Table 6, reversal position within the staircase series had little influence on reliability. Thus, the reliability coefficients of reversal combinations 2 + 3, 4 + 5 and 6 + 7 were 0.665, 0.705 and 0.673, respectively, suggesting that reversal position within the staircase had no meaningful impact on reliability. These combinations of reversals were chosen because they represented (i) averages of reversals from ascending and descending runs and (ii) successive thirds of the staircase series after the initial reversal. Inspection of the data presented in Table 6 also failed to show a systematic relationship between reliability and position within the reversal series beyond the low reliability of the first staircase

Table 5 Goodness-of-fit values (SSD and R^2) for the Spearman–Brown, linear, logarithmic and saturation models on data collected from eight olfactory tests. Values in bold correspond to the models depicted in Figure 1

Models	NP	Olfactory tests								
		UPSIT	MODSIT	Odor discrimination	Yes–No odor identification		Odor memory	PEA detection threshold	Intensity† rating (mean)	Pleasantness rating (mean)
					d′	Bias				
Spearman–Brown	1	0.0246	0.0227	**0.0100**	0 1143	0.1095	0.0184	**0.0018**	0.1130	0.0278
		77.3%	75.1%	**84.4%**	81.8%	66.0%	80.7%	**98.4%**	0.0%	13 9%
Linear	2	0.0195	0.0156	0.0099	0.1378	0.0575	**0.0049**	0.0181	**0.0018**	0.0021
		82.0%	82.9%	84.4%	78 1%	82.1%	**94.9%**	83.9%	**24.8%**	93 0%
Logarithmic	2	**0.0082**	0.0121	0.0097	0.1137	**0.0457**	0.0074	0.0032	0.0019	**0.0006**
		92.4%	86.7%	84.8%	81.9%	**85.8%**	92.2%	97.1%	23.3%	**98.2%**
Saturation models										
Spearman–Brown	2	0 0184	0.0227	0.0099	0.1007	0.0774	0 0184	0.0017	0.0024	0.0079
		83 0%	75.1%	84.4%	84.0%	76.0%	80.7%	98.5%	3.2%	75.5%
Linear	3	0.0090	**0.0070**	0.0100	**0.0624**	0.0562	0.0043	0.0080	0.0018	0.0010
		91.7%	**92.3%**	84.4%	**90.1%**	82.5%	95.5%	92.9%	32.5%	95.8%
Logarithmic	3	0.0072	0.0086	0.0097	0.1034	0.0457	0.0074	0 0026	0.0025	0.0006
		93.3%	90.6%	84.8%	83.6%	85 8%	92.2%	97.7%	23.3%	98.3%

SSD = sum of squared deviations or residuals, R^2 = percentage of variance accounted for; NP = the number of free parameters in the model

†Note that the R^2 values for this column are extremely low. These low R^2 values do not reflect poor fits of the model; in fact, the fit of the linear model is excellent (see Figure 1). Instead, the values are low because the R^2 statistic is highly sensitive to the variability of the dependent variable Particularly if the variance of the dependent variable is small, the R^2 value will be small even when the fit is almost perfect

reversal. As was expected (see Figure 1), (i) the first single staircase reversal value exhibited comparatively low reliability (r = 0.453), (ii) reliability increased, on average, as a function of the number of staircase reversals included in the analysis, and (iii) highest reliability was obtained by combining all of the reversal points (r = 0.845 for this data set).

Prediction of reliability of the UPSIT from the MODSIT

Since the Modular Smell Identification Test (MODSIT) was comprised of 12 UPSIT items, we applied the Spearman–Brown formula to see if the reliability of the full-length 40-item UPSIT was similarly predictable from a test 12/40ths as long. The predicted 40-item UPSIT value was 0.88, a value within 3% of the empirically-determined value of 0.90.

Discussion

The present data reveal that considerable variation is present in the reliability of olfactory tests administered in modern research and clinical settings. Such variation is problematic when inferences are made concerning underlying physiological processes of subjects and most likely explains discrepant findings among a number of studies in the literature. For example, Koss *et al.* (1988), in a study using 10 patients, concluded that olfactory detection (measured by the single ascending series butanol test; reliability = 0.47 in this study) and identification (measured by the UPSIT; reliability = 0.90 in this study) are dissociated in early Alzheimer's disease. Thus, a significant difference was observed between their AD patients and controls for the UPSIT, but not for the butanol threshold. However, when a staircase procedure is used for the threshold testing (reliability = 0.88 in this study), no such disparity between identification and detection test occurs (e.g. Doty *et al.*, 1987).

The reliability coefficients found in this experiment are of similar magnitude to the few reliability coefficients that have been reported in the literature. For example, we previously reported UPSIT test-retest and split-half reliability coefficients ranging from 0.87 to 0.95 (Doty *et al.*, 1984b, 1985, 1989); the 0.90 correlation observed in the present study is clearly within this range. Similarly, the test-retest

Table 6 Test-retest reliability coefficients (Pearson *r*'s) for segments of the PEA detection threshold test (seven total reversals collected)

Measure	*r*
Reversal #1	0.453
Reversal #2	0.595
Reversal #3	0.674
Reversal #4	0.647
Reversal #5	0.731
Reversal #6	0.641
Reversal #7	0.675
Reversal #1 + #2	0.657
Reversal #2 + #3	0.665
Reversal #3 + #4	0.764
Reversal #4 + #5	0.705
Reversal #5 + #6	0.767
Reversal #6 + #7	0.673
Reversal #1 + #2 + #3	0.692
Reversal #2 + #3 + #4	0 771
Reversal #3 + #4 + #5	0.774
Reversal #4 + #5 + #6	0.756
Reversal #5 + #6 + #7	0.753
Reversal #1 + #2 + #3 + #4	0.776
Reversal #2 + #3 + #4 + #5	0.793
Reversal #3 + #4 + #5 + #6	0.807
Reversal #4 + #5 + #6 + #7	0 762
Reversal #1 + #2 + #3 + #4 + #5	0.793
Reversal #2 + #3 + #4 + #5 + #6	0.827
Reversal #3 + #4 + #5 + #6 + #7	0.811
Reversal #1 + #2 + #3 + #4 + #5 + #6	0.832
Reversal #2 + #3 + #4 + #5 + #6 + #7	0.834
Reversal #1 + #2 + #3 + #4 + #5 + #6 + #7	0.845

reliability coefficient for butanol (r = 0.49) is similar to that noted by others. Cain and Gent (1991), for example, found, in a study of 32 subjects ranging in age from 22 to 59 years, that the correlation between butanol thresholds determined for the left and right sides of the nose (which they used as a reliability estimate), was, at best, 0.68 and as low as 0.30 when the butanol threshold test was the first in a series of four threshold tests. Punter (1983) reports test-retest reliability coefficients for butanol ranging from 0.15 to 0.42 in three separate tests of 31–38 subjects, although (i) the reliability coefficients determined in his study were routinely low for most of the odorants tested and (ii) a method of constant stimuli procedure was used, rather than the single ascending series used in this study. Heywood and Costanzo (1986) evaluated the test-retest reliability of the ascending butanol threshold procedure in 16 subjects aged 17–52 years. Test-retest reliability of the left nares was 0.45 (P = 0.08) and of the right nares was 0.08 (P = 0.76) (R. Costanzo, personal communication, June 12, 1995).

No relationship was found in this study between the reliability of any of the test measures and the duration of the test-retest periods. However, the data set was largely limited to test-retest intervals of less than 2 weeks. We previously reported that the short-term (≈2 weeks) test-retest reliability of the UPSIT was only marginally higher than the long-term (>6 months) test-retest reliability (respective r's = 0.92 and 0.89), implying that such a relationship for the UPSIT, if it exists, is weak (Doty *et al.*, 1984b, 1985).

A primary finding of the present work is that the reliability of olfactory tests is strongly related to their length. Importantly, of the three models evaluated, none optimally fit the reliability/test length data from all tests. Thus, the Spearman–Brown model provided the best fits to the PEA detection threshold and odor discrimination tests, whereas the linear saturation model fit the odor memory and yes/no identification d' data the best. The logarithmic model fit the UPSIT and mean pleasantness rating data the best, as well as the yes-no ID bias measure. In many of these cases, however, the closeness of fit of some of the models makes it difficult to be certain which model provides the absolutely best fit. However, the heterogeneity of fits among a number of models clearly suggests that test length and reliability are best described by different functions for different tests.

An argument can be made that the reliability of a test may be a more more important factor to consider in choosing an olfactory test for a specific application than the specific type of olfactory test (i.e. odor detection, identification, discrimination, etc.), particularly in cases where normal subjects are involved. This argument stems from the notion that nominally distinct olfactory tests may not, in many instances, measure dissimilar perceptual or physiological processes. Recently, Doty *et al.* (1994) administered nine of the ten olfactory tests evaluated in this paper to 97 subjects and performed a principal components analysis of the intercorrelation matrix. Four meaningful components emerged. The first was comprised of strong primary loadings from most of the olfactory test measures, whereas the second was comprised of primary loadings from intensity ratings given to a set of suprathreshold odorant concentrations. The third and fourth components had primary loadings that reflected, respectively, mean suprathreshold pleasantness ratings and the response bias measure derived from the yes-no odor identification test. Thus, for all practical purposes, most of the olfactory tests evaluated seem to measure a common source of variance, possibly akin to the 'G' factor observed in intelligence measurement theory (Spearman, 1904). Such a phenomenon could result if olfactory ability is largely determined by the degree of integrity of the olfactory epithelium, which undergoes considerable deterioration throughout the normal life span (see Curcio *et al.*, 1985; Nakashima *et al.*, 1984), presumably reflecting, in large degree, cumulative insults from viruses, airborne toxins and other environmental agents (Amoore, 1986; Deems *et al.*, 1991; Jiang *et al.*, 1974).

Clearly, future research is needed to better define to what degree nominally different olfactory tests actually tap different physiological processes. Until tests of high and equivalent reliability are administered to the same set of subjects, such definition will be enigmatic.

FOOTNOTES

1. Free parameters of a model give the model flexibility within a well-defined mathematical constraint. For example, a simple linear model has two free parameters: slope and *y*-incercept. The model is 'free', or unconstrained, with respect to how tilted the line is and where the line intersects the *y*-axis. However, the model is constrained to be a straight line; it is not allowed to have curvature. The more free parameters a model has, the more flexible it becomes, but also less falsifiable and, possibly, meaningful. A model that can fit well with few free parameters is preferable than one that fits the same data equally well using more free parameters. Note that, for the Spearman–Brown model, r_1 is the model's free parameter.
2. An approximate test can be constructed to assess the significance of the difference between two correlation coefficients when samples are not independent. For our application, the first correlation, r_{12}, is the test-retest reliability of, say, Test A, and is based on two measurements, M_1 and M_2 (i.e. test and retest); similarly, the second correlation, r_{34}, is the reliability of Test B based on measurements M_3 and M_4. Suppose all four measurements are taken on each of *n* subjects, and we wish to test whether r_{12} is significantly different from r_{34}. Thus, the null hypothesis is H_0: $\rho_{12} = \rho_{34}$, where r_{12} and r_{34} are the sample estimates of ρ_{12} and ρ_{34}, respectively. The test statistic is given by

$$Z = \frac{(r_{12}-r_{34})}{S},$$

where *S* is an estimate of the standard deviation of $r_{12} - r_{34}$. The statistic *Z* is distributed approximately normal with mean 0.0 and standard deviation 1.0; the usual hypothesis-testing procedure can be used to determine the statistical significance of the observed *Z*-score.

To obtain *S*, use

$$\sqrt{\mathrm{Var}\ (r_{12}-r_{34})} = \sqrt{\mathrm{Var}\ (r_{12}) + \mathrm{Var}\ (r_{34}) - 2\ \mathrm{Cov}\ (r_{12}, r_{34})}\,.$$

The approximations for Var (r_{12}) and Var (r_{34}) are given in Anderson (1984, Theorem 4.2.4.) and are:

$$\mathrm{Var}\ (r_{12}) \approx \frac{1}{n-1}\ (1-r_{12}^2)^2,$$

$$\mathrm{Var}\ (r_{34}) \approx \frac{1}{n-1}\ (1-r_{34}^2)^2,$$

The approximation for Cov(r_{12},r_{34}) is determined by extending Anderson's technique and is given by

$$\mathrm{Cov}\ (r_{12}, r_{34}) \approx \frac{1}{n-1}[r_{13}r_{24} + r_{14}r_{23} - r_{12}\ (r_{13}r_{14} + r_{23}r_{24}) - r_{34}(r_{13}r_{23} + r_{14}r_{24}) + \frac{1}{2}\ r_{12}r_{34}(r_{13}^2 + r_{14}^2 + r_{23}^2 + r_{24}^2)].$$

3. Although overall test scores were available for all 57 subjects upon which to compute test-retest reliability coefficients (Table 3), in some cases specific item or reversal data were missing from the data set. Therefore, the *n*'s upon which the data in Figure 1 and Table 5 were calculated were: UPSIT—55; MODSIT—51; Odor discrimination test—53; Yes–No odor Identification Test—56 (for both *d′* and bias measures); Odor memory test—50; Odor intensity and pleasantness rating tests—55. For the PEA detection threshold test, 15 subjects were omitted from the modeling analyses who did not provide a total of seven reversals on either the first or second test occasion (e.g. as a result of ceiling or floor effects).
4. The determination of the best-fitting model was made without the benefit of statistical comparisons. Our criteria for selecting the best-fitting model were as follows: (i) if two models have the same number of free parameters, the model with the smaller SSD was selected as the better

model; and (ii) if two models do not have the same number of free parameters, but have the same or similar SSDs (i.e. difference between two SSDs is less than 0.001), the model with fewer free parameters was selected as the better model since, by chance alone, models with more free parameters will produce better fits.

ACKNOWLEDGEMENTS

Supported by grants PO1 DC 00161 from the National Institute on Deafness and Other Communication Disorders, and RO1 AG 08148 from the National Institute of Aging, National Institutes of Health.

REFERENCES

Amoore, J.E. (1986) Effects of chemical exposure on olfaction in humans. In Burrow, C.S. (ed.), *Toxicology of the Nasal Passages*, Hemisphere Publishing, Washington, DC, pp. 155–190.

Amoore, J.E. and Ollman, B.G (1983) Practical test kits for quantitatively evaluating the sense of smell. *Rhinology*, **21**, 49–54.

Anderson, T.W. (1984) *An Introduction to Multivariate Statistical Analysis*, 2nd edn. Wiley, New York

Cain, W.S. and Gent, J.F. (1991) Olfactory sensitivity. reliability, generality and association with aging. *J. Exp. Psychol.: Hum. Percept. Perform.*, **17**, 382–391.

Cain, W.S. and Rabin, R.D. (1989) Comparability of two tests of olfactory functioning. *Chem. Senses*, **14**, 479–485.

Cain, W.S., Gent, J.F., Goodspeed, R.B. and Leonard, G. (1988) Evaluation of olfactory dysfunction in the Connecticut Chemosensory Clinical Research Center. *Laryngoscope*, **98**, 83–88.

Cain, W.S., Cometto-Muniz, J.E. and De Wijk, R.A. (1992) Techniques in the quantitative study of human olfaction. In Serby, M.J. & Chobor, K L (eds), *Science of Olfaction*. Springer-Verlag, New York. pp. 279–308.

Chapman, L.J. and Chapman, J P. (1978) The measurement of differential deficit. *J. Psychiat. Res.*, **14**, 303–311.

Corwin, J. (1989) Olfactory identification in hemodialysis: acute and chronic effects on discrimination and response bias. *Neuropsychologia*, **27**, 513–522.

Cronbach, L.J. (1960) *Essentials of Psychological Testing*. Harper & Row, New York.

Curcio, C A., McNelly, N.A. and Hinds, J.W. (1985) Aging in the rat olfactory system: relative stability of piriform cortex contrasts with changes in olfactory bulb and olfactory epithelium. *J. Comp. Neurol.*, **235**, 519–528.

Deems, D.A. and Doty, R.L. (1987) Age-related changes in the phenyl ethyl alcohol odor detection threshold. *Trans. Penn. Acad. Ophthalmol. Otolaryngol.*, **39**, 646–650.

Deems, D.A., Doty, R.L., Settle, R.G., Moore-Gillon, V., Shaman, P, Mester, A.F., Kimmelman, C.P., Brightman, V.J. and Snow, J.B., Jr (1991) Smell and taste disorders: a study of 750 patients from the University of Pennsylvania Smell and Taste Center (1981–1986) *Arch. Otolaryngol. Head Neck Surg.*, **117**, 519–528.

Doty, R.L. (1975) An examination of relationships between the pleasantness, intensity and concentration of 10 odorous stimuli *Percept Psychophys.*, **17**, 492–496.

Doty, R.L. (1991) Olfactory system. In Getchell, T.V., Doty, R.L., Bartoshuk, L.M. & Snow, J.B. Jr (eds), *Smell and Taste in Health and Disease*. Raven Press, New York, pp. 175–203.

Doty, R.L (1992) Diagnostic tests and assessment. *J. Head Trauma*, **7**, 47–65.

Doty, R.L. (1995) *Handbook of Olfaction and Gustation*. Marcel Dekker, New York.

Doty, R.L., Frye, R.E. and Agrawal, U. (1989) Internal consistency reliability of the fractionated and whole University of Pennsylvania Smell Identification Test. *Percept. Psychophys.*, **45**, 381–384.

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.

Doty, R.L., Shaman, P., Applebaum, S.L., Giberson, R., Sikorsky, L. and Rosenberg, L. (1984a) Smell identification ability: changes with age. *Science*, **226**, 1441–1443.

Doty, R.L., Shaman, P. and Dann, M. (1984b) Development of the University of Pennsylvania Smell Identification Test: a standardized microencapsulated test of olfactory function (Monograph). *Physiol. Behav.*, **32**, 489–502.

Doty, R.L., Newhouse, M.G. and Azzalina, J.D. (1985) Internal consistency and short-term test-retest reliability of the University of Pennsylvania Smell Identification Test. *Chem. Senses*, **10**, 297–300.

Doty, R.L., Reyes, P. and Gregor, T. (1987) Presence of both odor identification and detection deficits in Alzheimer's disease. *Brain Res. Bull.*, **18**, 597–600.

Doty, R.L., Smith, R., McKeown, D.A. and Raj, J (1994) Tests of human olfactory function: principle components analysis suggests that most measure a common source of variance *Percept. Psychophys.*, **56**, 701–707.

Doty, R.L., Marcus, A. and Lee, W.W. (1995) Development of the 12-Item Cross-Cultural Smell Identification Test. *Laryngoscope*, in press.

Guilford, J.P. (1954) *Psychometric Methods*. McGraw-Hill, New York.

Harper, R., Bate-Smith, E.C. and Land, D.G. (1981) *Odour Description and Odour Classification*. American Elsevier, New York.

Heywood, P.G and Costanzo, R.M. (1986) Identifying normosmics: a comparison of two populations. *Am. J. Otolaryngol.*, **7**, 194–199.

Jiang, X.Z., Buckley, L.A. and Morgan, K.T. (1974) Pathology of toxic responses to the RD50 concentration of chlorine gas in the nasal passages of rats and mice. *Toxicol. Appl. Pharmacol.*, **71**, 225–236.

Koss, E , Weiffenbach, J.M., Haxby, J.V. and Friedland, R P. (1988) Olfactory detection and identification performance are dissociated in early Alzheimer's disease. *Neurology*, **38**, 1228–1232.

Köster, E P. (1975) Human psychophysics in olfaction. In Moulton, D.G., Turk, A. and Johnston, J.W., Jr (eds), *Methods in Olfactory Research*. Academic Press, New York, pp. 345–374.

Magnusson, D. (1967) *Test Theory*. Addision-Wesley, Reading, MA.

Nakashima, T., Kimmelman, C.P. and Snow, J.B., Jr (1984) Structure of human fetal and adult olfactory neuroepithelium. *Arch. Otolaryngol.*, **110**, 641–646.

Proetz, A.W. (1924) Exact olfactometry. *Annl. Otol. Rhinol. Laryngol.*, **33**, 275–278.

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

Smith, R.S., Doty, R.L., Burlingame, G.K. and McKeown, D.A. (1993) Smell and taste function in the visually impaired. *Percept. Psychophys.*, **54**, 649–655.

Snodgrass, J.C. and Corwin, J. (1988) Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *J. Exp. Psychol.: Gen.*, **117**, 34–50.

Spearman, C. (1904) 'General Intelligence', objectively determined and measured. *Am. J. Psychol.*, **15**, 201–293.

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

Takagi, S.F. (1989) *Human Olfaction*. University of Tokyo Press, Tokyo.

Toulouse, E. and Vaschide, N. (1899) Mesure de l'odorat chez l'homme et chez la femme. *Comp. Rend. Soc. Biol.*, **51**, 381–383.

Valentin, G (1848) *Lehrbuch der Physiologie des Menschen*. Braunschweig.

Wenzel, B. (1948) Techniques in olfactometry. *Psychol. Bull.*, **45**, 231–246.

Wilkinson, L. (1990) *SYSTAT: The System for Statistics*. SYSTAT, Inc., Evanston, IL.

Yoshida, M. (1984) Correlation analysis of detection threshold data for 'standard test' odors *Bull. Facul. Sci. Eng. Chuo Univ.*, **27**, 343–353.

Zwaardemaker, H. (1889) On measurement of the sense of smell in clinical examination. *Lancet* I, 1300–1302.

Received on March 30, 1995; accepted on June 26, 1995