C S

'Sniffin' Sticks': Olfactory Performance Assessed by the Combined Testing of Odor Identification, Odor Discrimination and Olfactory Threshold

T. Hummel[1,4], B. Sekinger[1], S.R. Wolf[2], E. Pauli[3] and G. Kobal[1]

[1]Department of Experimental and Clinical Pharmacology and Toxicology, University of Erlangen-Nürnberg, Krankenhausstraße 9, 91054 Erlangen, [2]Department of Oto-Rhino-Laryngology, University of Erlangen-Nürnberg, Waldstraße 1, 91054 Erlangen, [3]Department of Neurology, University of Erlangen-Nürnberg, Schwabachanlage 6, 91054 Erlangen, Germany and [4]Smell and Taste Center, Department of Otorhinolaryngology, University of Pennsylvania, 3400 Spruce Street, Philadelphia, PA 19104, USA

Correspondence to be sent to: Dr Thomas Hummel, Smell and Taste Center, Department of Otorhinolaryngology, University of Pennsylvania, 3400 Spruce Street, Philadelphia, PA 19104, USA

Abstract

'Sniffin' Sticks' is a new test of nasal chemosensory performance based on pen-like odor dispensing devices. It comprises three tests of olfactory function, namely tests for odor threshold (*n*-butanol, testing by means of a single staircase), odor discrimination (16 pairs of odorants, triple forced choice) and odor identification (16 common odorants, multiple forced choice from four verbal items per test odorant). After extensive preliminary investigations the tests were applied to a group of 104 healthy volunteers (52 female, 52 male, mean age 49.5 years, range 18–84 years) in order to establish test–retest reliability and to compare them with an established measure of olfactory performance (the Connecticut Chemosensory Clinical Research Center Test, CCCRC). Performance decreased with increasing age of the subjects ($P < 0.001$). Coefficients of correlation between sessions 1 and 2 were 0.61 for thresholds, 0.54 for discrimination and 0.73 for identification. Butanol thresholds as obtained with the CCCRC increased as a function of age; this relation to the subjects' age was not found for the CCCRC odor identification task. The test–retest reliability for CCCRC thresholds was 0.36, for odor identification it was 0.60. It is concluded that 'Sniffin' Sticks' may be suited for the routine clinical assessment of olfactory performance. **Chem Senses 22: 39–52, 1997.**

Introduction

Tests for the assessment of olfactory function are numerous. However, in the clinical practice of otorhinolaryngology or neurology few, if any, of them are actually used. The reasons may be found in the inconsistency of some tests, the lack of normative data, the time needed for administration and the limited availability of these tests (for review see Doty and Kobal, 1995). This situation very often reduces clinical testing of olfactory ability to the administration of one or two common odors (e.g. coffee or cloves) in combination with the question whether the patient is able to identify the odor. The lack of an appropriate means for the testing of olfactory function limits the quality of medical

diagnosis; it also prevents quality control in the treatment of disorders associated with the sense of smell.

In North America the so-called UPSIT (University of Pennsylvania Smell Identification Test; Doty, 1989) or its down-scaled version, the CC-SIT (Cross Cultural Smell Identification Test; Doty *et al.*, 1996) has reached the widest degree of distribution. The UPSIT is a 'scratch and sniff' test based on microencapsulated odorants which are released from the surface of strips by means of a pencil. The patient is asked to tag 40 odorants on a multiple-choice list comprising four items each. Since its introduction in 1984 it has been thoroughly investigated (for review see Doty and Kobal, 1995). Cain and Rabin (1989) combined threshold testing with an odor identification task in the Connecticut Chemosensory Clinical Research Center Test (CCCRC). In that test, the threshold of subjects for butanol is assessed by means of squeeze-bottles using the method of ascending limits; odor identification is performed by means of eight bottles containing different odorants with a multiple choice from a list of 16 items identical for all odorants. Amoore (1992) proposes various test kits based on squeeze bottles for special purposes, e.g. evaluation of hyposmia or hyperosmia, specific anosmias, odor recognition, discrimination and description.

In Europe and Asia (except Japan) no such test has experienced a supra-regional distribution. In addition, as indicated above, available tests either focus on specific olfactory tasks such as verbal odor identification (e.g. the UPSIT), are not available on a commercial basis (e.g. the CCCRC) or may not appeal to the budget-oriented clinician (see 'Arbeitsgemeinschaft Olfaktologie und Gustologie' of the German Society for Otorhinolaryngology, 1994). The aim of the present study was to create a re-usable and portable test-kit of olfaction which would include both verbal (odor identification) and non-verbal approaches (odor discrimination) in combination with an elaborate odor threshold testing. In addition, the test should utilize the subjects' sniffing behaviour (Laing, 1983) rather than the administration of squeeze bottles, which always produces tactile sensations as a result of headspace-air being squeezed from the bottle. Finally, the test should utilize cost-effective materials to make it attractive for a clinician to use.

Accordingly, in the recent study, built on the concept of pen-like odor dispensing devices, smell containing felt-tip pens were used that received publicity in grammar schools in the seventies. When sealed with an appropriate cap olfactory contamination of the environment is eliminated and the drying out of the pen is prevented. When opened the felt-tip ensures that odorants are presented in a constant concentration similar to the delivery of dye when writing. The naming of these devices as 'Sniffin' Sticks' was inspired by a remark of Dr Donald A. Leopold, Johns Hopkins Hospital, Baltimore, MD; in the following the test will often be referred to as Sticks.

In parallel to the olfactory test presented in this paper the 'Arbeitsgemeinschaft Olfaktologie und Gustologie' of the German Society for Otorhinolaryngology (1994) also encouraged the development of a simple screening test of olfactory function which is described elsewhere (Kobal *et al.*, 1997).

Materials and methods

All experimental procedures were explained and demonstrated in full detail to the subjects, who provided written informed consent. The study was performed in accordance with the Declaration of Helsinki/Hong Kong. In the following, materials and methods are presented as far as they were an integral part of most experiments performed. Specific details of the individual experiments are presented together with the results.

'Sniffin' Sticks'

Odorants were presented in commercially available (unfilled) felt-tip pens. The pens had a length of ~14 cm, with an inner diameter of 1.3 cm. Instead of liquid dye the tampon was filled with liquid odorants or odorants dissolved in propylene glycol, to a total volume of 4 ml. For odor presentation the cap was removed by the experimenter for ~3 s and the pen's tip was placed ~2 cm in front of both nostrils. Possible bacterial contamination of the Sticks was checked regularly over a period of 4 months. In no instance was the presence of pathogenic microorganisms found.

Odor identification

According to others (Cain and Rabin, 1989; Doty, 1989), odor identification should be assessed by means of common odors. Twenty-one odorants were chosen, the aim being to select 16 for inclusion in the final version of the 'Sniffin' Sticks'. The relatively small number of 16 odorants was chosen in order to satisfy the time restrictions encountered by clinicians when testing olfactory performance. Criteria for the selection of odorants were as follows: (i) subjects should be familiar with all odor-describing items used in the test; (ii) odorants included in the test should be similar with

regard to both intensity and hedonic tone; and (iii) the successful identification of individual odorants from a list of four descriptors should be >75% in healthy subjects. Subjects were free to sample the odors as often as necessary to make a decision. Each odorant was presented by the experimenter and there was an interval of at least 30 s to prevent olfactory desensitization (Hummel *et al.*, 1996).

Odor discrimination

Odor discrimination was performed by means of triplets of odorants. The subject was presented with three odorants and the task was to identify the sample that had a different smell (Kobal *et al.*, 1992). To prevent visual detection of the target sticks, subjects were blindfolded with a sleeping mask. As with odor identification, only 16 triplets were to be selected in order to meet clinical time restrictions for testing. Criteria for the selection of odorants were similar as described above for odor identification: (i) odorants in a triplet should be similar with regard to intensity; it also appeared to be of advantage if odors were similar in their hedonic tone; and (ii) correct discrimination of individual odorants should be >75% in healthy subjects. To keep the time needed for testing to a minimum, other than with the odor identification task subjects were only once allowed to sample the odor. Presentation of triplets was separated by at least 30 s. The interval between presentation of individual sticks of a triplet was ~3 s.

Odor thresholds

Odor thresholds were assessed using *n*-butanol as the odorant (Cain and Rabin, 1989); dilutions were established in a geometric series (Cain and Rabin, 1989; Kobal *et al.*, 1992). Preliminary experiments aimed to establish the appropriate dilution difference between stimuli. Presentation of the odorants was similar to that described above for the discrimination task. Subjects were blindfolded to prevent visual identification of some of the odorant-containing sticks. Using a triple-forced-choice paradigm, detection thresholds were determined by employing a single staircase method as described by Doty (1991). Three sticks were presented to each subject in a randomized order, two contained the solvent and the other the odorant at a particular dilution. The task of the subject was to indicate the stick with the odorant. Presentation of the triplets to a subject occurred every 20 s, until they had correctly discerned the odorant in two successive trials which triggered a reversal of the staircase. The geometric mean of the last four staircase reversal points of a total of seven reversals (Doty, 1991) was used as the threshold estimate. The duration of this procedure varied between 10 and 20 min.

The Connecticut Chemosensory Clinical Research Center Test

The CCCRC consists of two tests of olfactory performance: an odor identification task and the determination of the *n*-butanol threshold. Presentation of odorants for threshold testing is performed by means of squeeze bottles (squeezable bottles made from plastic) while odor identifcation is assessed by means of sniff bottles (bottles made from glass). The highest concentration of butanol in the series was 4% in water; 11 successive dilutions were established as a geometric series dilution ratio of 1:3. Testing was performed with the concentrations in ascending series using a two-alternative, forced choice paradigm by which patients have to identify the odorant containing bottle after both odorant and blank have been administered (double-alternative, forced choice paradigm). The threshold was defined as the concentration at which subjects succeed to identify *n*-butanol in five successive trials. The odor identification task employed eight items (baby powder, chocolate, cinnamon, coffee, mothballs, peanut butter, ivory soap and Vicks Vaposteam). Patients were given a list with 16 terms comprising eight terms describing the items used in the test and eight items describing other common items.

All odorants were handled most carefully; the experimenters always wore deodorized disposable cotton gloves. Measurements were performed in quiet, well-ventilated rooms.

Acoustic rhinometry

As a measure of nasal congestion, the volume of the anterior part of the nasal cavity was assessed by acoustic rhinometry (Rhinoklack®, Stimotron, Germany). Volumetric measures were taken over a distance of 3 cm starting from the nasal valve (Roithmann *et al.*, 1994; Min and Jang, 1995).

Statistical analyses

Results were analyzed by means of the SPSS/PC+ program package as follows: the data were submitted to ANOVAs (analysis of variance, repeated measurements design), and intercorrelations were computed between experimental variables.

Results

Preliminary experiments

Odor identification

Identification of odorants was performed as a multiple forced choice from a list of four descriptors. During preliminary experiments one aim was to establish that the target items used in the list were familiar to most of the subjects. This was done by means of questionnaires. Subjects rated the familiarity of 92 pre-selected items on a numerical scale where '1' meant 'highly familiar' and '6' 'unknown'. Only those items were included into the multiple choice that had ratings of '1' or '2' in >90% of the 63 subjects investigated (32 male, 31 female; mean age 29.3 years).

In addition, to ensure homogeneity of the test only odorants were included which were similar to each other in terms of identifiability and intensity. Both identifiability and intensity were rated on visual analog scales. Thus, 21 pre-selected odorants were roughly matched for odor intensity; they were then tested with a group of 27 healthy subjects (14 male, 13 female; mean age 26 years). As a criterion set by the authors, only those odorants were selected for the final version of the odor identification task that had a familiarity rating of ≥89%, a rating of identifiabilty ±25% of the mean identifiability and an intensity rating ±25% of the mean perceived intensity (see Table 1). In terms of the hedonic tone of the odorants care was taken to include more odorants with a pleasant smell; this was done in an effort to make the test more acceptable to subjects or patients. The ratings of the odorants' hedonic tone exhibited a good correlation (r_{16} = 0.79) with data established by Dravnieks *et al.* (1984).

Regarding odor identification, it was also determined whether additional non-verbal information would improve the subjects' ability to identify odorants. A typical example of the combined administration of verbal and non-verbal information is presented in Figure 1. Twenty-five subjects

Table 1 Characterization of the 16 odorants selected for the identification task [results of an investigation in 32 male and 31 female subjects (mean age 29 years)]

Odorant	Familiarity	Identifiability	Intensity	Hedonics	Dravnieks
Orange	95	–8	–13	24	2.86
Peppermint	100	14	13	25	2.50
Turpentine	92	–8	–3	–20	–0.73
Cloves	94	–3	4	10	1.67
Leather	97	–13	–23	–10	1.30
Banana	100	6	9	32	2.00
Garlic	97	11	16	–17	–0.17
Rose	93	–17	–18	14	3.08
Fish	100	6	9	–36	–1.98
Lemon	100	–3	3	21	2.50
Coffee	100	1	–3	–1	2.33
Anise	90	12	10	35	1.21
Cinnamon	96	3	0	46	2.54
Liquorice	89	12	–5	–7	1.21
Apple	95	–9	0	18	2.61
Pineapple	97	–5	3	20	2.59

Familiarity was rated on a 6-point numerical scale ranging from 1 = 'highly familiar' to 6 = 'unknown'. The number indicates the percentage of subjects who rated the item with a score of ≤4. *Identifiability* ratings were made on a visual analog scale of 10 cm length (left-hand end defined as not identifiable = 0 units, right-hand end defined as absolutely identifiable = 100 units). The numbers indicate the percentage deviation from the mean identifiability of these 16 odorants (mean = 85.2 units, SD 8.3 units). *Intensity*: ratings of perceived intensity were made on a visual analog scale of 10 cm length (left-hand end defined 'no odor perceived' = 0 units, right-hand end defined as 'highest intensity possible' = 100 units). The numbers indicate the percent of deviation from the mean perceived intensity of these 16 odorants (mean = 77.2 units, SD 8.4 units). *Hedonics*: hedonic tone was rated on a visual analog scale of 10 cm length (left-hand end defined as absolutely unpleasant = –50 units, right-hand end defined as absolutely pleasant = +50 units). The mean rating of the 16 odorants was +9.5 units (SD 22.7 units). *Dravnieks*: the numbers in this column refer to data provided by Dravnieks *et al.* (1984); the more positive the number the more pleasant was the hedonic estimate (absolutely unpleasant = –3.5, absolutely pleasant = +3.5). The coefficient obtained for the correlation between the hedonic tone rated by the subjects and the data provided by Dravnieks *et al.* was r_{16} = 0.79.

(13 male, 12 female; mean age 24 years) participated in this study, in which 21 odorants had to be identified using a multiple choice task; 13 subjects received both verbal and non-verbal information, 12 subjects received only the list of verbal items. The non-verbal information did not improve the percentage of correctly identified odorants as compared to the sole verbal presentation of the items (verbal and non-verbal information: identification rate 85%; verbal information: identification rate 86%). Results from that experiment were also used to select the 16 odorants to be included into the final set (see Table 1). Only odorants were included which had an identification rate ≥80%. Odorants that could not be included were menthol, onion, lilac, melon and raspberry. The coefficient of correlation between mean ratings of identifiability and the percentage of correctly identified odorants was $r_{25} = 0.67$ ($P < 0.001$).

Odor discrimination

For odor discrimination, triplets of odorants were presented. To prevent subjects from discriminating properties of odorants other than quality, the two odorants of each triplet had to be balanced in terms of their intensity. Thus, after the concentrations for the 20 pairs of odorants had been established, they were investigated in 21 subjects (10 male, 11 female; mean age 25 years). As criterion selected by the authors only those pairs of odorants were permitted which had a maximum mean intensity difference of ≥25%. Only those pairs were chosen that could be discriminated by ~75% of the subjects (15/21; Table 2). For the enantiomers of carvone a lesser degree of discriminability was allowed; they were included to widen the range of the possible difficulty of the individual tasks.

Odor thresholds

The following experiments were performed in 25 subjects (12 male, 13 female, mean age 24 years) and aimed to identify which dilution ratio of *n*-butanol (1:2 or 1:3 respectively; highest concentration 4%) was better suited for

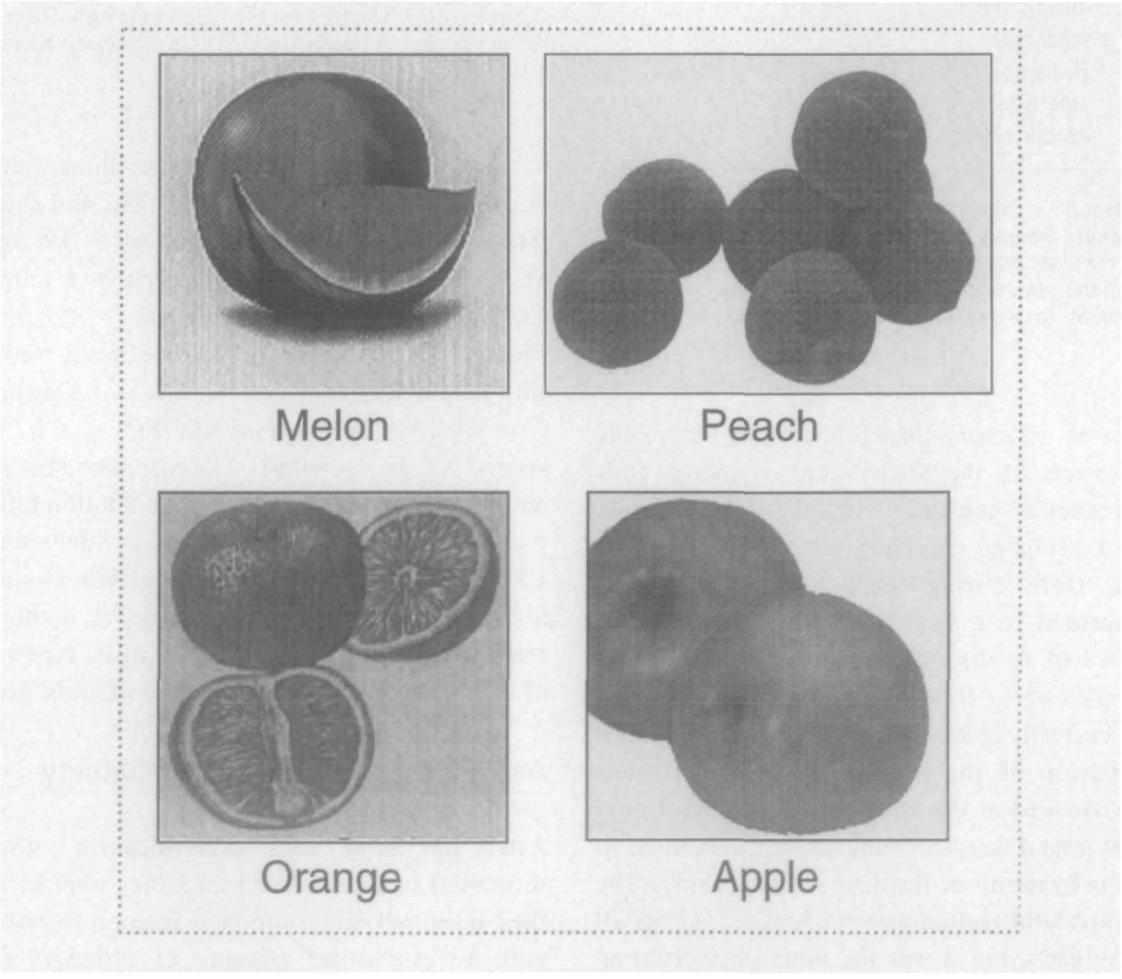

Figure 1 Example of a card presenting both verbal and non-verbal information used for odor identification. The addition of non-verbal information did not improve the subjects' performance in odor identification.

Table 2 Characterization of the 16 pairs of odorants selected for the discrimination task [results of an investigation in 10 male and 11 female subjects (mean age 25 years)]

Target odorants	Non-target odorants	Difference in intensity	% correctly identified
			+
Butanol	2-phenyl ethanol	5	71
Isoamylacetate	anethole	17	90
Anethol	eugenol	17	90
Limonene	fenchone	22	86
(–) Carvone	(+) carvone	1	43
Eugenol	cinnamon aldehyde	6	90
Dihydro rosenoxide	menthol	5	76
Acetaldehyde	isoamylacetate	4	81
Citronellal	linalool	8	76
Pyridine	limonene	7	81
Limonene	citronellal	22	95
Eucalyptol	dipyridyl	18	86
Dipyridyl	cyclopenta-decanoate	9	71
Butanol	fenchone	3	86
Octylacetate	cinnamon aldehyde	10	81
Carvone	acetaldehyde	4	81

Subjects had to identify the target stimuli. *Difference in intensity*: differences in perceived intensity of two odorants of a pair are presented in arbitrary units; the mean difference was 9.9 units (SD 6.9 units). *% correctly identified*: the percentage of subjects who correctly identified the inidividual odorant. The mean identification rate was 80.3% (SD 12.2 units).

the assessment of olfactory thresholds. Thus, thresholds obtained by means of the Sticks were compared with *n*-butanol thresholds assessed with the CCCRC test (dilution ratio 1:3; highest concentration 4%; see Cain and Rabin, 1989). Only during these measurements was *n*-butanol presented in a triple-forced choice paradigm where subjects had to discriminate *n*-butanol from two blanks. Thresholds were defined as the concentration where *n*-butanol was correctly identified four times in a row. As an additional criterion of the test's quality the test–retest reliability was assessed at the same time on different days separated by at least 4 days. All subjects were submitted to threshold testing by means of the three different assays. The sequence of threshold testing was randomized across all participating subjects, but it was the same when retesting individual subjects.

The experiments revealed that *n*-butanol concentrations

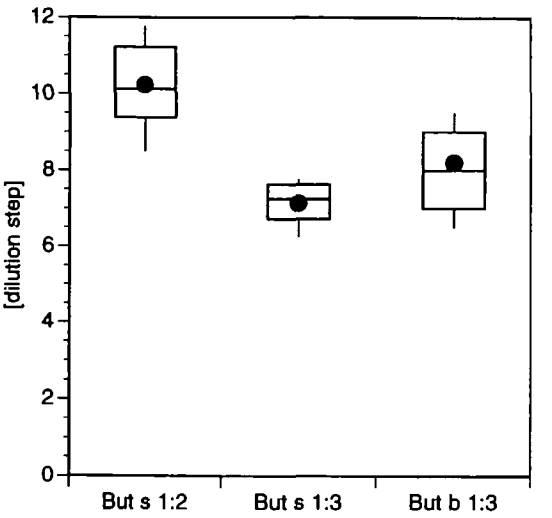

Figure 2 Box plots of threshold measurements in 25 subjects for *n*-butanol presented either in squeeze bottles (dilution ratio 1:3—But b 1:3) or in Sticks (dilution ratio 1:2—But s 1:2; or 1:3—But s 1:3) respectively. Each set of data relates to a box plot; the lowest, second lowest, middle, second highest and highest point represents the 10, 25, 50, 75 and 90 percentile respectively. Means are displayed as filled circles. Butanol concentrations necessary to produce threshold sensations were approximately in the same range for both sticks and squeeze bottles [Sticks 1:2: mean 10.2 dilution steps (= 1.6 µg butanol/l), Sticks 1:3: mean 7.1 dilution steps (= 8.1 µg butanol/l); CCCRC: mean 8.2 dilution steps (= 1.9 µg butanol/l)].

necessary to produce threshold sensations were in approximately the same range for both sticks and squeeze bottles [Sticks 1:2, mean 10.2 dilution steps (= 1.6 µg butanol/l); Sticks 1:3, mean 7.1 dilution steps (= 8.1 µg butanol/l); CCCRC, mean 8.2 dilution steps (= 1.9 µg butanol/l); Figure 2]. Correlation between the mean results obtained with the three assays was good (Sticks 1:3 versus Sticks 1:2, r_{25} = 0.92; Sticks 1:3 versus CCCRC, r_{25} = 0.78; Sticks 1:2 versus CCCRC, r_{25} = 0.66). The correlation between test and retest was highest for the Sticks at a dilution ratio of 1:2 (r_{25} = 0.49, $P < 0.05$). Based on this result, Sticks at a dilution of 1:2 were chosen for further testing. When compared with Sticks at a dilution of 1:3, Sticks at a dilution of 1:2 exhibited a larger variation, indicating its superior capability of resolving small differences in the subjects' performance.

Assessment of test–retest reliability, comparison with CCCRC

After the three tests (identification, discrimination, threshold) had been established they were investigated for their retest reliability, and their relation to results obtained with an established measure of olfactory function. In addition, the results were investigated in relation to both age and gender of the subjects. The experimenter rated the

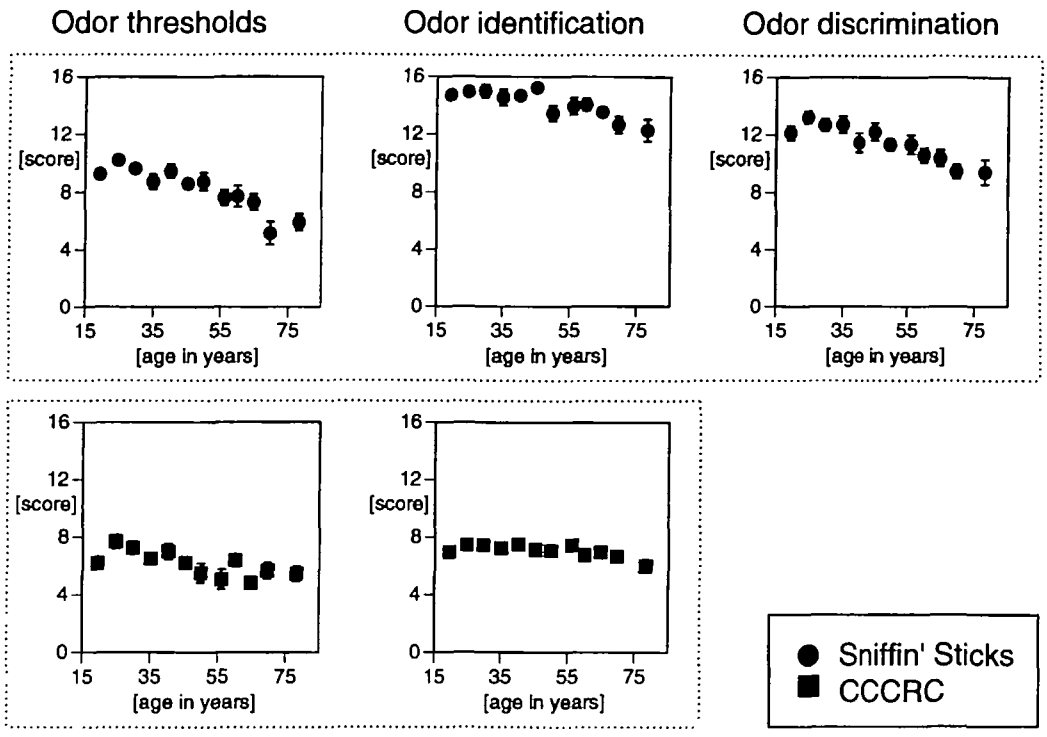

Figure 3 Mean results for both Sticks and CCCRC ($8 \leq n \leq 11$, means, SEM) in relation to the subjects' age. A significant effect of the factor 'age' was observed by analysis of variance for all three tests of the Sticks ($P < 0.001$) but only for the CCCRC's threshold test ($P < 0.05$). The mean decrease of performance was most pronounced in subjects older than 60 years (see Table 3).

subjects' educational background in three categories. They received a score between 1 and 3 [1 = subjects who graduated from college ('Hauptschule') or blue collar workers; 2 = subjects with a high-school degree ('mittlere Reife') or white collar workers; 3 = subjects with a university degree or a position in academia or management]. Tests were carried out on a total of 104 Caucasian subjects (52 male and 52 female; mean age 49.5 years, SD 18.5; age range 18–84 years); 54 of these subjects were tested in the morning, 52 in the afternoon. The mean interval between test and retest was 10 days (SD 11 days). Assessment of olfactory performance by means of either the Sticks or the CCCRC test was separated by an interval of ~60 min; in half of the subjects (26 male and 26 female) testing was performed first with the CCCRC followed by testing with the Sticks. In addition, in one half of the subjects measurements were performed for the left nostril, while in the other half of the subjects they were performed for the right nostril.

For statistical analyses 12 groups were defined based on the subjects' age (group 1: 18–22 years; group 2: 23–27 years; group 3: 28–32 years; group 4 33–37 years; group 5: 38–42 years; group 6: 43–47 years; group 7: 48–52 years; group 8: 53–57 years; group 9: 58–62 years; group 10: 63–67 years; group 11: 68–72 years; group 12: 73–84 years). There were eight subjects in each group with the exception of groups 10 ($n = 11$), 11 ($n = 10$) and 12 ($n = 11$). Groups were balanced for the number of male or female subjects, the nostril tested and the assay with which testing started (i.e. CCCRC or Sticks). Results were analysed by means of MANOVAs with factors 'age', 'sex' or 'left/right' as between-subject factors and 'session' as the within-subject factor.

Sticks

A significant effect of the factor 'age' was observed for odor identification [$F(92/11) = 4.02$, $P < 0.001$], odor discrimination [$F(92/11) = 4.66$, $P < 0.001$] and for butanol thresholds [$F(92/11) = 8.42$, $P < 0.001$]. For all measurements performance decreased with increasing age of the subjects (Figure 3) and this decrease was most pronounced in subjects older than 65 years. In addition, the factor 'session' was always statistically significant for both odor identification ($F \geq 8.62$, $P < 0.01$) and odor thresholds ($F \geq 5.61$, $P < 0.05$), indicating an increase in the subjects' performance in session 2. No significant effects of the factor 'session' were observed for odor discrimination (Figure 4); discrimination was smaller during the first session when the left nostril was stimulated than when the right nostril was

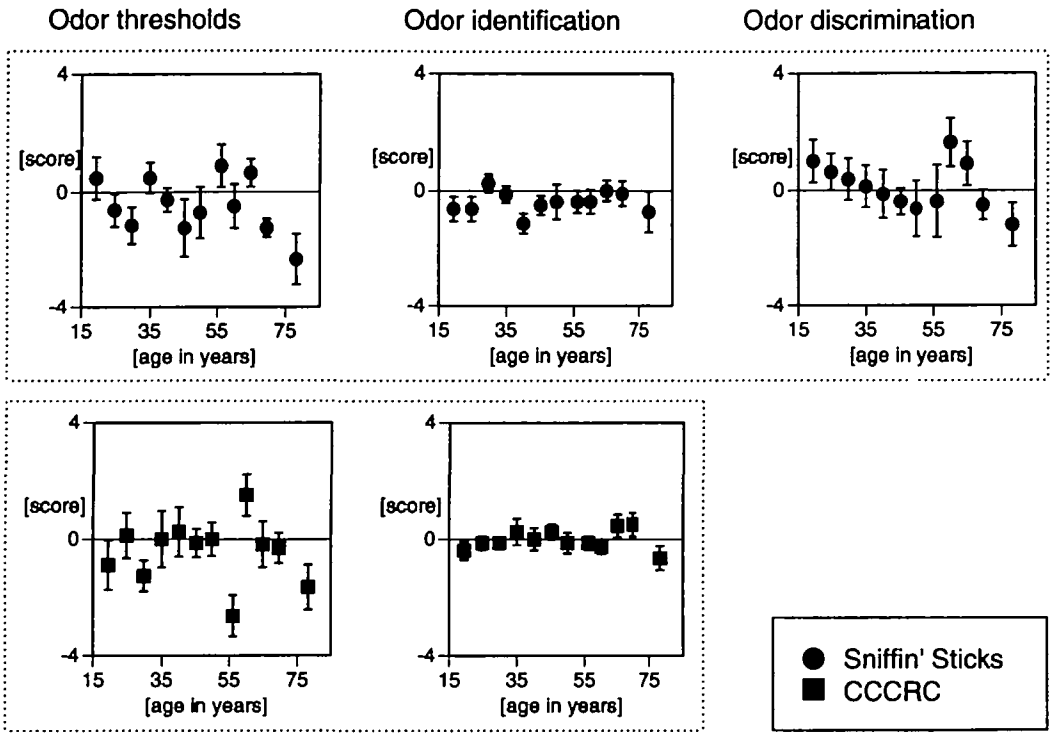

Figure 4 Mean differences between results obtained in sessions 1 and 2 (retest) for both Sticks and the CCCRC in relation to the subjects' age ($8 \leq n \leq 11$, means, SEM). Negative figures indicate an increase of the subjects' score when olfactory performance was retested. The factor 'session' was significant for the Sticks' identification and threshold task and for the CCCRCs threshold measurements ($P < 0.05$), indicating a relative increase of the subjects' performance in session 2; this difference between test and retest was more pronounced in older subjects than in younger subjects (see Table 3).

stimulated; this difference between the two nostrils was much larger than in session 2 (Figure 5). The differences became significant as an interaction between factors 'left/right' and 'session' [$F(102/1) = 4.65$, $P < 0.05$]. For odor thresholds there was an additional interaction between factors 'age' and 'session' [$F(92/12) = 2.14$, $P < 0.05$, indicating that the increase of the subjects' performance from session 1 to session 2 was the more pronounced the older the subjects (Figure 4).

CCCRC

When analysing results of the CCCRC only thresholds decreased as a function of age [factor 'age', $F(91/11) = 3.14$, $P < 0.05$] but not the subjects' performance in the identification task [factor 'age', $F(91/11) = 1.81$, $P < 0.1$; Figure 3]. As with the Sticks in the odor threshold task, an interaction was found between the factors 'age' and 'session' [$F(92/11) = 2.01$, $P < 0.05$]. That is, during session 1 odor thresholds were higher than those obtained in session 2; this was most pronounced in elderly subjects (Figure 4).

Correlations between measurements

The coefficients of correlation between results obtained

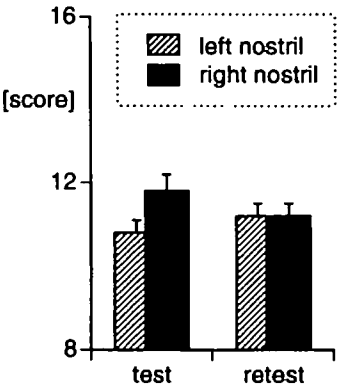

Figure 5 Odor discrimination scores (means, SEM, $n = 52$) in relation to both stimulation of the left or right nostril and the session when testing was performed. The significant interaction between factors 'side' and 'session' indicated that the right nostril advantage disappeared when retested.

during the first and second sessions were $r_{104} = 0.61$ for odor thresholds, $r_{104} = 0.54$ for odor discrimination and $r_{104} = 0.73$ for odor identification. Coefficients of correlations between olfactory performance and the subjects' educational background were $0.01 < r_{104} < 0.04$.

When a composite score was computed for the three olfactory tests (mean of the percentage of the three scores in

relation to their respective mean; Table 3) the coefficient of correlation between test and retest was $r_{104} = 0.72$. This coefficient was slightly diminished when only two of the three tasks were used for computation of the composite score, i.e. odor thresholds and odor discrimination ($r_{104} = 0.66$), odor thresholds and odor identification ($r_{104} = 0.71$) or odor identification and odor discrimination ($r_{104} = 0.68$). For the composite score the coefficient of correlation with the subjects' age was $r_{104} = 0.69$; it decreased to $r_{104} = 0.68$ when either scores of the threshold measurements and discrimination task or threshold scores and identification task were analysed. Correlation to the subject's age was smallest ($r_{104} = 0.61$) when scores of odor identification and odor discrimination were used to compute the composite score.

Table 3 Composite score of olfactory performance derived from the three tests of the 'Sniffin' Sticks' (mean ± SEM, $8 \leq n \leq 11$)

Mean age	Sessions 1 + 2	Session 1	Session 2
19	109.2 ± 1.8	111.0 ± 3.5	107.5 ± 2.1
25	116.9 ± 1.9	115.8 ± 3.1	118.0 ± 1.9
30	113.2 ± 2.2	111.7 ± 1.6	114.8 ± 3.0
35	108.4 ± 4.4	109.4 ± 4.9	107.4 ± 4.5
40	108.0 ± 3.5	105.9 ± 4.6	110.1 ± 3.4
45	108.0 ± 2.5	104.2 ± 3.8	111.7 ± 3.7
50	101.7 ± 3.8	98.8 ± 4.9	104.5 ± 4.4
56	98.3 ± 4.8	99.1 ± 5.5	97.4 ± 5.6
60	96.7 ± 4.7	97.6 ± 4.6	95.8 ± 5.7
65	93.2 ± 3.0	95.9 ± 3.6	90.6 ± 3.4
70	79.4 ± 5.1	76.0 ± 5.8	82.9 ± 4.7
78	81.5 ± 5.9	74.0 ± 6.7	89.0 ± 6.1

The score was defined as the mean of the subjects' performance in the three individual tests after they had been expressed as a percentage of the average score reached by the 104 volunteers tested.

Regarding the thresholds as assessed by means of the CCCRC, the coefficient of correlation between test and retest was $r_{104} = 0.36$; for odor identification it was $r_{104} = 0.60$. Figure 6 gives an overview of the correlations between different tests obtained for the mean performance averaged across test and retest; the statistical significance of all intercorrelations listed was $P \leq 0.001$.

Correlations between the subject's age and volume of the anterior nasal cavity indicated that the subject's nasal cavities became wider with increasing age (Figure 7). This correlation was larger for subjects where only the right nostril had been tested than for subjects where the left nostril was tested (Table 4). An additional finding was that

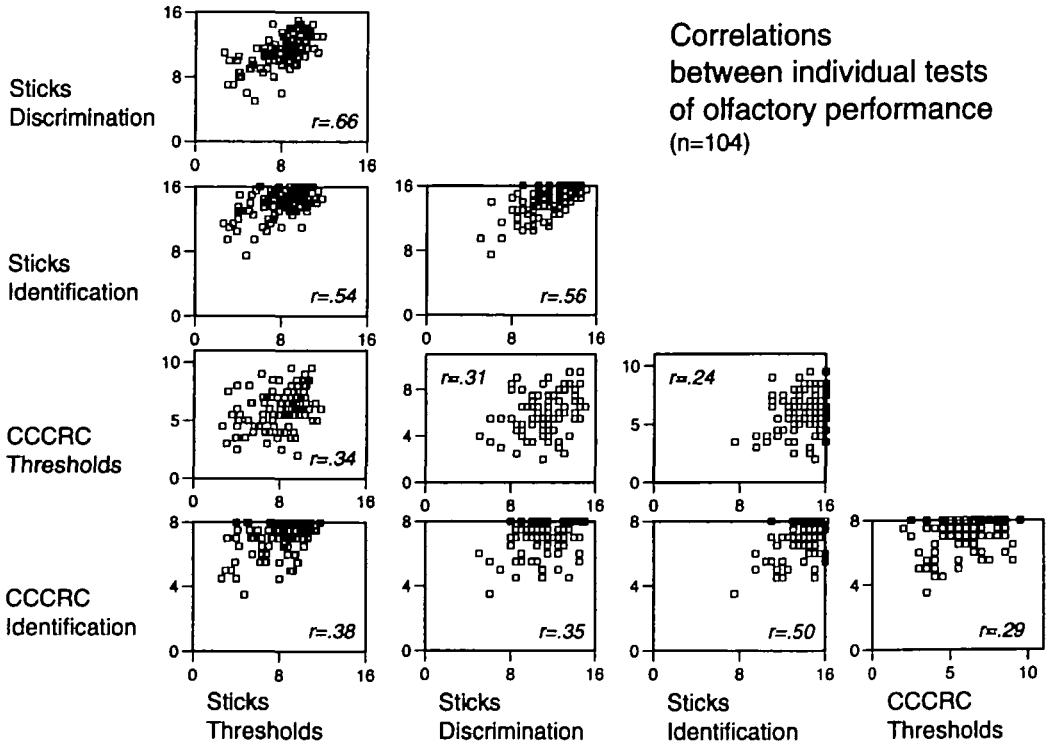

Figure 6 Intercorrelations between different tests of olfactory performance ($n = 104$). The y-axes of individual graphs are adjusted to the maximum range of scores; coefficients of correlation are indicated with italic letters.

correlations were relatively smaller for the side where testing had been performed compared with the other side of the nose. When correlating nasal volume with olfactory performance, a trend towards a negative relation emerged, suggesting that olfactory function would decrease with an increase in the volume of the nasal cavity. However, when correlations were computed separately for subjects below and above the age of 50, the results indicated that this trend towards negative coefficients of correlations was mostly due to elderly subjects (Table 5).

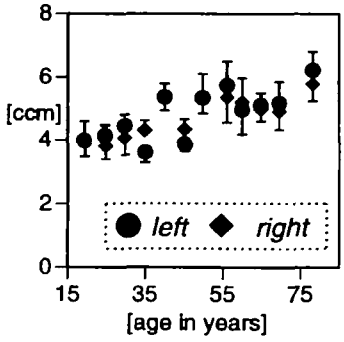

Figure 7 Volume of anterior nasal cavity in relation to the subjects' age measured separately for the left and the right nostril (means, SEM; $8 \leq n \leq 11$). Volumes tended to exhibit a symmetrical, bilateral increase across life-span.

Table 4 Coefficients of correlations between age and volume of anterior nasal cavities as assessed by means of acoustic rhinometry

	Left nasal cavity	Right nasal cavity
Left nostril tested	$r_{51} = 0.24$	$r_{51} = 0.25$
Right nostril tested	$r_{53} = 0.47$	$r_{53} = 0.37$

Long-term testing

Repetitive testing over a period of 4 months was performed in six subjects (two male, four female, mean age 27 years). These tests were initiated in order to investigate the long-term reproducibility of olfactory performance; in addition, the stability of the test kit needed to be investigated on a functional level.

Intra-individual testing was always performed at the same time of day; in half of the subjects measurements were performed for the left nostril, in the other half for the right nostril. Tests were performed on a total of seven occasions after preparation of the kit, i.e. on average subjects were tested after 8, 24, 43, 63, 79, 98 and 131 days. None of the three measures of olfactory performance exhibited a major difference between mean results obtained in the seven consecutive measurements (Figure 8). The mean coefficients of correlation between the day of testing and the scores in the three olfactory tasks averaged across the six subjects were <0.1 (threshold: $r_6 = -0.01$; discrimination: $r_6 = 0.08$; identification: $r_6 = -0.04$). In contrast, the anterior nasal volume exhibited a slight increase over time ($r_6 = 0.48$).

Discussion

The experiments described above indicate that the 'Sniffin' Sticks' are suited for olfactory testing. The coefficient of correlation between test and retest was 0.73 for odor identification, 0.61 for odor thresholds and 0.54 for odor discrimination. Similar data have been reported by Doty *et al.* (1995), who found the highest test–retest reliability for odor identification tasks and the lowest one for odor discrimination. The reproducibility of results obtained with the Sticks was much higher than our results obtained with

Table 5 Coefficients of correlations between olfactory performance and volume of anterior nasal cavities in relation to the subjects' age

	Age 18–84 years: nostril tested		Age 18–49 years: nostril tested		Age 50–84 years: nostril tested	
	Left (n = 52)	Right (n = 52)	Left (n = 27)	Right (n = 25)	Left (n = 24)	Right (n = 28)
Sticks: thresholds	–0.24	–0.23	0.23	–0.03	–0.04	–0.07
Sticks: discrimination	–0.26	–0.25	–0.21	–0.01	–0.03	–0.12
Sticks: identification	–0.20	–0.10	0.05	0.06	–0.18	0.16
CCCRC: thresholds	–0.17	–0.39	–0.22	–0.08	0.06	–0.38
CCCRC: identification	–0.20	–0.07	0.24	0.25	–0.27	–0.03

the CCCRC, from which coefficients of correlations were 0.36 for odor thresholds and 0.60 for odor identification respectively. Cain and Gent (1991) reported a coefficient of correlation of $r_{32} = 0.68$ for thresholds obtained for the left and right nostril as an estimate of test–retest reliability. Earlier, Cain and Rabin (1989) observed a coefficient of correlation between results obtained for the left and right nostril to be as high as $r_{50} = 0.90$. For the correlation between tests performed on two different days separated by ~2 weeks, other researchers have found a coefficient of $r_{57} = 0.49$ for butanol thresholds (Doty *et al.*, 1995).

Other tests, based on odor identification tasks only, have produced higher coefficients of correlation between test and retest. For the GITU ('Geur Identificatie Test Utrecht', 36 odorants) and the UPSIT ('University of Pennsylvania Smell Identification Test', 40 odorants) these coefficients were 0.96 and 0.95 respectively (Hendriks, 1992). This good test–retest reliability may be due to the use of larger numbers of odorants than the 16 used in the present study for odor identification. For example, when only 12 odorants of the UPSIT are used, the coefficient of correlation between test retest decreases to 0.71 (Doty *et al.*, 1995).

Compared with the CCCRC, the Sticks were better suited for determining the life-span decrease of olfactory function, which has been established in numerous studies (Venstrom and Amoore, 1968; Schiffman and Pasternak, 1979; Cain and Gent, 1991; Ship and Weiffenbach, 1993; Murphy *et al.*, 1994; T. Hummel *et al.*, submitted for publication). The present data also confirm the findings of Doty *et al.* (1984) that a strong decrease of the subjects' ability to identify odorants occurred above the age of 65 years. The present findings extend those of Doty *et al.* in that a corresponding decrease of olfactory function was also observed in the same subjects for odor discrimination and odor thresholds (compare Deems and Doty, 1987; Schiffmann and Pasternak, 1979). A single cause of the pronounced decrease of performance beyond the age of 60–65 years has not yet been identified; aside from psychological factors such as age-related changes in attention or memory, it is likely that it is based on a number of factors, predominantly epithelial

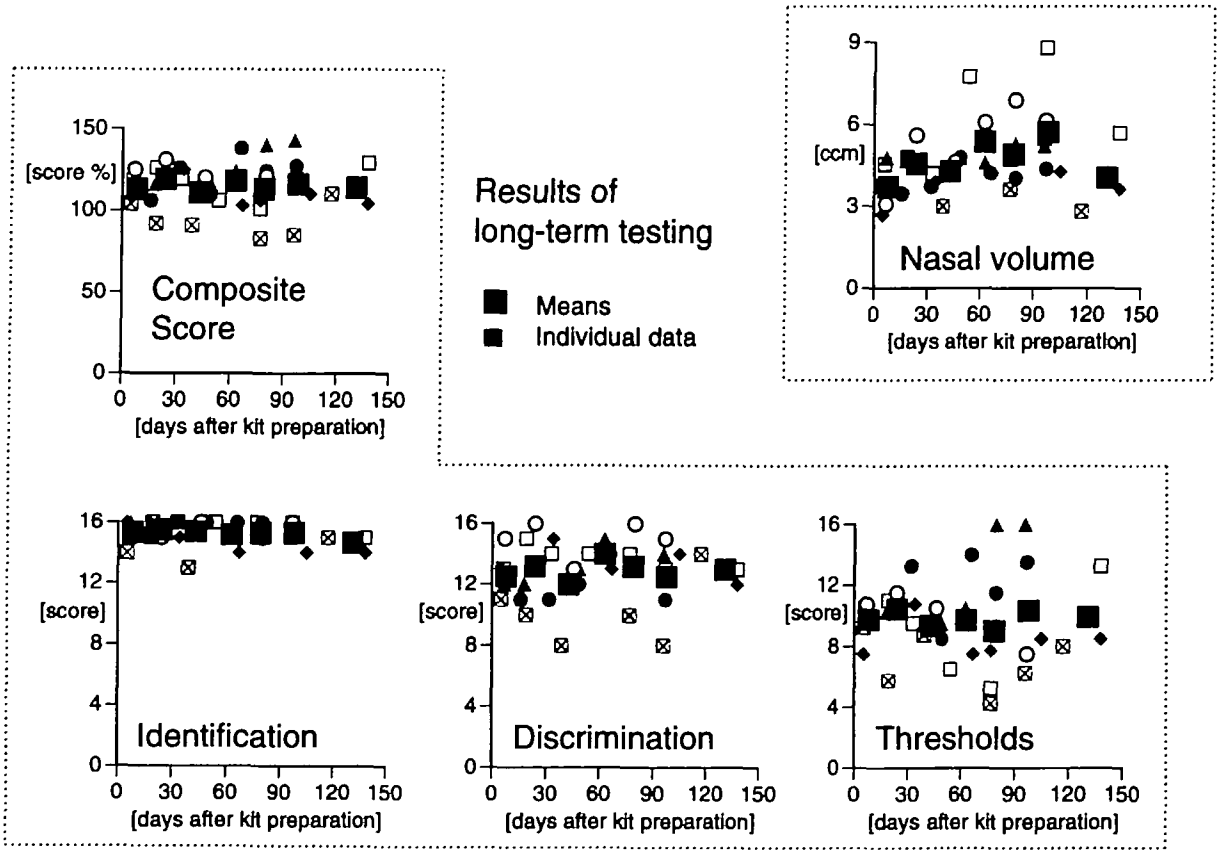

Figure 8 Results of long-term testing in relation to the day when the kit had been prepared. Means across the six subjects are presented as filled, black squares; individual data are shown as smaller symbols. The variance of data was greatly reduced when a composite score was derived from the three olfactory tests in relation to the mean performance in the respective task as previously established in 104 subjects (see above). Mean olfactory performance exhibited little variation over time; in contrast, the volume of the anterior nasal cavity increased slightly over the observation period of ~4 months.

alterations, e.g. reduced metabolism, occlusion of the cribriform plate, changes of epithelial blood flow and increased mucus viscosity (for review see Doty, 1994).

Odor discrimination was significantly better during the first session when the right nostril was tested than when the left nostril was tested; this difference between the two nostrils disappeared in the second session. Zatorre and Jones-Gotman (1990) reported a right nostril advantage in an odor discrimination task (eight pairs of odorants) that was investigated in a total of 99 healthy volunteers who participated in one session. The authors related their findings to a relative specialization of function within the right cerebral hemisphere. The present findings extend these previous reports in showing that learning processes seem to play a more specific role in discrimination tasks than all other tests administered where subjects' did always significantly better during the second session. This specific involvement of memory functions in odor discrimination may explain the relatively low coefficient of correlation between test and retest ($r_{104} = 0.54$).

As indicated by the significant interaction between the factors 'session' and 'age' for thresholds assessed with either Sticks or the CCCRC, elderly subjects exhibited lower thresholds during the second session than during session one. Since this phenomenon was only found for olfactory thresholds and not for odor identification or discrimination, this suggests a training effect specifically related to odor thresholds.

When building a composite score of all three tests of olfactory function, the coefficient of correlation between test and retest was $r_{104} = 0.72$. This coefficient was slightly diminished when only two of the three tasks were used for

Table 6 Coefficients of correlations between test and retest when different reversals of the staircase were used for computation of thresholds

Staircase reversals used for computation of threshold score	Coefficient of correlation between test and retest (n = 104)	Rank of coefficient of correlation	Rank of coefficient of correlation from Doty *et al.* (1995)
1	0.20	28	27
2	0.26	27	26
3	0.46	21	20
4	0.47	16	24
5	0.59	6	17
6	0.60	5	25
7	0.41	24	21
1 + 2	0.27	26	23
2 + 3	0.42	23	22
3 + 4	0.46	20	14
4 + 5	0.55	12	19
5 + 6	0.46	19	13
6 + 7	0.56	10	20
1 + 2 + 3	0.40	25	19
2 + 3 + 4	0.46	18	11
3 + 4 + 5	0.50	15	10
4 + 5 + 6	0.57	8	15
5 + 6 + 7	0.63	1	16
1 + 2 + 3 + 4	0.44	22	9
2 + 3 + 4 + 5	0.54	13	8
3 + 4 + 5 + 6	0.56	9	6
4 + 5 + 6 + 7	0.61	2	12
1 + 2 + 3 + 4 + 5	0.46	17	7
2 + 3 + 4 + 5 + 6	0.55	11	4
3 + 4 + 5 + 6 + 7	0.60	4	5
1 + 2 + 3 + 4 + 5 + 6	0.51	14	3
2 + 3 + 4 + 5 + 6 + 7	0.60	3	2
1 + 2 + 3 + 4 + 5 + 6 + 7	0.57	7	1

the composite score, i.e. odor thresholds and odor discrimination (r_{104} = 0.66), odor thresholds and odor identification (r_{104} = 0.71) or odor identification and odor discrimination (r_{104} = 0.68). These findings indicate that a composite score may be better suited for the clinical assessment of olfactory dysfunction than an isolated measure of olfactory performance such as odor identification, as suggested by Cain and Rabin (1989). The present data suggest that it makes sense to test olfactory performance using olfactory tests in different combinations. That is, when conducting non-verbal olfactory testing a combination of odor thresholds and odor discrimination appears to produce results of similar reliability to the combined determination of odor thresholds and odor identification, which might be appropriate in patients without cognitive deficits.

In the present study thresholds were assessed by averaging the last four of seven turning points, as reported by Doty (1991). Only recently, a paper from the same laboratory indicated that the test–retest reliability of phenyl ethyl alcohol threshold measurements increases when more than the last four turning points are included in the threshold measure (Doty *et al.*, 1995). However, when analysing our threshold data in a similar fashion (Table 6), the use of the last three turning points produced the highest coefficient of correlation between test and retest (r_{104} = 0.63) followed by the 'traditional' approach using the last four turning points (r_{104} = 0.61). Reasons for this discrepancy may be found either in the different concentration steps or in the different odorants that were used with the different tests. Doty used a logarithmic scale while the present investigation employed dilutions with a ratio of 1:2. Regarding the Sticks, the higher resolution of small differences in thresholds resulted in a larger variance in the responses than the method used by Doty. In other words, in the determination of odor thresholds, because of the lower resolution, large steps between separate concentrations may result in a very constant performance and thus in a high test–retest reliability which by itself does not tell much about the quality of a test. It is conceivable that with 'high-resolution' methods thresholds are best determined from the last turning points of a staircase where the subject is thoroughly acquainted with both the method and the setting.

Nasal volume increased as a function of age. This finding may be considered as a result of decreased intranasal blood flow (Bende, 1983), decreased responsiveness to autonomic stimulation (Hasegawa and Kern, 1977) or age-related atrophy (Somlyo and Somlyo, 1968; Nishihata, 1984; Murphy *et al.*, 1985). Thus, an increase of nasal patency is not always related to increased olfactory sensitivity.

The three olfactory tests did not exhibit a major change when tested over 4 months. In contrast to measures of olfactory function, the volume of the anterior nasal cavity appeared to increase over time. It appears likely that the seasons might have affected the congestional state of the nasal mucosa, as measurements started in late August and were finished by November. In turn, the present data argue against a major influence of small changes in the congestional state of the nasal mucosa on olfactory function, which extends previous research on changes in olfactory function in relation to major changes of airflow dynamics (e.g. Rehn, 1978; Schwartz *et al.*, 1987; Doty and Frye, 1989; Delank, 1992).

The present data indicate that the Sticks may prove a useful tool for the clinical testing of the nasal chemical senses. Within the framework of a multinational collaborative effort the test is currently being applied in more than 10 centers in Germany, Switzerland, Austria and Italy. Since all the centers have agreed to report results to the authors of this paper, the database of the Sticks is expected to expand continuously, which in turn will strengthen the diagnostic power of the 'Sniffin' Sticks'.

REFERENCES

'Arbeitsgemeinschaft Olfaktologie und Gustologie' of the German Society of Otorhinolaryngology (1994) *Protokoll im Rahmen des Deutschen HNO-Kongresses am 16.5.1994 in Chemnitz, Germany.*

Amoore, J.E. (1992) Odor standards in squeeze-bottle kits for matching quality and intensity. *Water Sci. Tech.*, **25**, 1–9.

Bende, M. (1983) Bloodflow with ^{133}Xe in human nasal mucosa in relation to age, sex, and body position. *Acta Otolaryngol. (Stockh.)*, **96**, 175–179.

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

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

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

Delank, K.W. (1992) Die olfaktorische Sensitivität bei der Rachenmandelhyperplasie. *Laryngo-Rhino-Otol.*, **71**, 293–297.

Doty, R.L. (1989) *The Smell Idenfication Test® Administration Manual*, 2nd edn. Sensonics, Haddonfield.

Doty, R.L. (1991) Olfactory system. In Getchell, T.V., Doty, R.L., Bartoshuk, L.M. and Snow, J.B., Jr (eds), *Smell and Taste in Health and Disease*. Raven Press, New York, p. 1803.

Doty, R.L. (1994) Olfactory dysfunction in the elderly and in Alzheimer's disease. In Kurihara, K., Suzuki, N. and Ogawa, H. (eds), *Olfaction and Taste XI*. Springer-Verlag, Tokyo, pp. 597–601.

Doty, R.L. and Frye, R. (1989) Influence of nasal obstruction on smell function. *Otolaryngol. Clin. North Am.*, **22**, 397–411.

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., Sikorski, L. and Rosenberg, L. (1984) Smell identification ability: changes with age. *Science*, **226**, 1441–1443.

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

Doty, R.L, Marcus, A. and Lee, W.W. (1996) Development of the 12-item cross-cultural smell identification test (CC-SIT). *Laryngoscope*, **106**, 353–356.

Dravnieks, A., Masurat, T. and Lamm, R.A. (1984) Hedonics of odors and odor descriptors. *J. Air Poll. Control Assoc.*, **34**, 752–755.

Hasegawa, M. and Kern, B. (1977) The human nasal cycle. *Mayo Clin. Proc.*, **52**, 28–34.

Hendriks, A.P.J. (1992) Gitu versus UPSIT. A comparison between two clinical tests of odour identification. *Chem. Senses*, **17**, 838–839.

Hummel, T., Knecht, M. and Kobal, G. (1996) Peripherally obtained electrophysiological responses to olfactory stimulation in man: electro-olfactograms exhibit a smaller degree of desensitization compared with subjective intensity estimates. *Brain Res.*, **717**, 160–164.

Kobal, G., Barz, S. and Hummel, T. (1992) A combined psychophysical and electrophysiological olfaction test. *Chem. Senses*, **17**, 850–851.

Kobal G., Hummel T., Sekinger B., Barz S., Roscher S. and Wolf S.R. (1997) 'Sniffin' Sticks': a screening test of olfactory performance. *Acta Oto-Rhino-Laryngol.*, in press.

Laing, D.G. (1983) Natural sniffing gives optimum odor perception for humans. *Perception*, **12**, 99–107.

Min, Y.G. and Jang, Y.J. (1995) Measurements of cross-sectional area of the nasal cavity by acoustic rhinometry and CT scanning. *Laryngoscope*, **105**, 757–759.

Murphy, C., Nunez, K., Whithee, J. and Jalowayski, A.A. (1985) The effects of age, nasal airway resistance and nasal cytology on olfactory threshold for butanol. *Chem. Senses*, **10**, 418.

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

Nishihata, S. (1984) Aging effects in nasal resistance. *Nippon Jibiinkoka Gakkai Kaiho*, **87**, 1654–1671.

Rehn, T. (1978) Perceived odor intensity as a function of air flow through the nose. *Sens. Proc.*, **2**, 198–205.

Roithmann, R., Cole, P., Chapnik, J., Barreto, S.M., Szalai, J.P. and Zamel, N. (1994) Acoustic rhinometry, rhinomanometry, and the sensation of nasal patency: a correlative study. *J. Otolaryngol.*, **23**, 454–458.

Schiffman, S. and Pasternak, M. (1979) Decreased discrimination of food odors in the elderly. *J. Gerontol.*, **34**, 73–79.

Schwartz, D.N., Mozell, M.M., Youngentob, S.L., Leopold, D.L. and Sheehe, P.R. (1987) Improvement of olfaction in laryngectomized patients with the larynx bypass. *Laryngoscope*, **97**, 1280–1286.

Ship, J.A. and Weiffenbach, J.M. (1993) Age, gender, medical treatment, and medication effects on smell identification. *J. Gerontol.*, **48**, M26–M32.

Somlyo, A.P. and Somlyo, A.V. (1968) Vascular smooth muscle I. Normal structure, pathology, biochemistry, and biophysics. *Pharmacol. Rev.*, **20**, 197–272.

Venstrom, D. and Amoore, J.E. (1968) Olfactory threshold in relation to age, sex, or smoking. *J. Food Sci.*, **33**, 264–265.

Zatorre, R.J. and Jones-Gotman, M. (1990) Right-nostril advantage for discrimination of odors. *Percept. Psychophys.*, **47**, 526–531.

Received on April 23, 1996; accepted on October 7, 1996