Internal consistency and short-term test-retest reliability of the University of Pennsylvania Smell Identification Test

Richard L. Doty¹,²,³, Marisa G. Newhouse¹,³ and Jeffrey D. Azzalina⁴
¹Smell and Taste Center, ²Department of Physiology, ³Department of Otorhinolaryngology and Human Communication, School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, and ⁴Sensonics, Inc., Philadelphia, PA 19143, USA

Abstract. Measures of internal consistency and short-term test-retest reliability of the University of Pennsylvania Smell Identification Test (UPSIT) were determined in 69 people who showed a broad range of test scores at an earlier test administration. In addition, two procedures for releasing the microencapsulated odorants were compared. The split-half reliability coefficients for the two administrations were \( r = 0.930 \) and \( r = 0.957 \), respectively (\( p < 0.001 \)). The test-retest (~ 2-week interval) coefficient was \( r = 0.949 \) (\( p < 0.001 \)). Release of the odorants by the tip of a #2 lead pencil resulted in scores equivalent to those obtained by releasing them with #120 sandpaper. These data, along with those collected previously on the long-term test-retest reliability (\( r = 0.918 \)), indicate that the UPSIT is a highly reliable and internally consistent test of smell function.

Introduction

The development of the University of Pennsylvania Smell Identification Test (UPSIT) now makes it possible for researchers and clinicians to accurately and conveniently measure smell function without the use of complex olfactometric equipment or cumbersome sniff bottles (Doty et al., 1984a,b). Because this test can be self-administered and sent through the mail, it is amenable to applications outside the traditional laboratory setting. In addition, norms based upon scores from over 1600 subjects are currently available for this test (Doty, 1983).

Although the long-term test-retest reliability (6 months) of this measuring instrument is known (\( r = 0.918 \); Doty et al., 1984a), there has been no assessment of its internal consistency or short-term test-retest reliability. On numerous theoretical grounds, short-term reliability is generally expected to exceed long-term reliability (Cronbach, 1960). Since the UPSIT correlates \( r = 0.91 \) with cerebral spinal fluid levels of a major neuro-metabolite of norepinephrine in certain patient groups (Mair et al., 1983), its short-term reliability coefficient likely exceeds \( r = 0.91 \), since the correlation of a test with a criterion measure theoretically can only be equal to or less than its reliability measure.

The primary purpose of this study was to determine the split-half and the short-term test-retest reliability of the UPSIT. Such measures provide information on the internal consistency and self-correlation of a test, as well as its stability and dependability over time. A secondary purpose was to establish whether UPSIT scores obtained by releasing the microencapsulated odorants with a pencil tip differ from those obtained by releasing them with #120 sandpaper. Earlier work using 9-point category scales noted that different release methods subtly influenced rated intensity and other psychological dimensions of microencapsulated odorants, although it was not determined if scores in a forced-choice identification task are similarly influenced (Doty et al., 1984a).

Methods

Materials
The UPSIT is discussed in detail in a number of other publications (Doty, 1983; Doty et al., 1984a). Briefly, this standardized test consists of four booklets containing 10 odorants apiece, one odorant per page. The stimuli are embedded in 10 - 50 μm diameter microcapsules fixed in a proprietary binder and positioned on brown strips at the bottom of each page. Above each odorant strip is a multiple-choice question with four response alternatives for each item. The subject is required to answer one of the four alternatives, even if no smell is perceived (i.e., the test is forced-choice). The stimuli span a wide range of qualitative odor classes and are composed of both single and multiple component odorants (e.g., cinnamon odor is represented by the single-component substance, cinnamaldehyde, whereas lime odor is represented by the complex natural essence of lime). Details of the stimuli used in this test are presented elsewhere (Doty et al., 1984a).

Subjects
Thirty five men (mean age = 53.23, SD = 16.93) and thirty-four women (mean age = 46.76, SD = 17.81) were selected for this study from a large subject pool maintained by the Smell and Taste Center. Fifty-eight of these persons were Caucasian Americans, nine were Black Americans, and one was an Asian Indian. These subjects had taken the UPSIT a year or more before the present testing, and were chosen to represent, in approximately equal numbers, the following intervals of initial UPSIT test scores: 10 - 15, 16 - 20, 21 - 25, 26 - 30, 31 - 35 and 36 - 40. These intervals were chosen to ensure that an adequate sampling of the entire range of test scores was made so that valid correlations could be computed.

Procedures
The subjects were administered the UPSIT on two occasions separated from one another by ~2 weeks. On the first occasion, half were instructed to release the stimuli with the tip of a #2 pencil, and the other half were instructed to use a 4.8 x 2.4 cm strip of #120 sandpaper. The order of the release procedure was reversed for the second test. The subjects were paid $15 apiece for their participation.

Results
The split-half (odd-even) reliability coefficient was determined for each of the two test administrations following correction for test length using the Spearman-Brown formula (Cronbach, 1960). These values were $r = 0.930 (p < 0.001)$ and $r = 0.957 (p < 0.001)$, respectively.

The short-term test-retest reliability was $r = 0.95 (p < 0.001)$; Figure 1) for the entire group, $r = 0.939 (p < 0.001)$ for the men and $r = 0.964 (p < 0.001)$ for the women. The mean test scores for each of the release procedures were as follows. Sandpaper: males = 26.9 (SD = 10.46), females = 28.2 (SD = 10.42); pencil: males = 26.4 (SD = 11.52), females = 27.9 (SD = 10.06). Although the average scores of the women were slightly higher than those of the men, this was not significant (gender by release
Consistency and reliability of UPSIT

procedure ANOVA: $F(1,65) = 0.31, p > 0.20$). Test order (sandpaper first versus pencil first) and release procedure (sandpaper versus pencil) were also not significant (order $F(1,65) = 0.77, p > 0.20$; release procedure $F(1,65) = 0.95, p > 0.20$).

Discussion

The present data clearly demonstrate that the UPSIT is a highly reliable and internally consistent measure of the ability to identify odors. Furthermore, they indicate that in a forced-choice identification task the release of the microencapsulated stimuli with the tip of a #2 lead pencil or by the surface of #120 sandpaper results in equivalent test scores.

The reliability of tests of olfactory function are rarely reported in the literature. Jones (1955) found the reliability of recognition thresholds for n-butanol, safrol and n-butyric acid (using sniff bottles in a method of limits paradigm) to be reasonably high (respective intra-class $r$ values = 0.82, 0.77 and 0.80). Although Punter (1983) found group thresholds for 11 compounds and relatively large numbers of subjects were highly reliable across two measurement sessions (Pearson $r = 0.95$), individual test-retest thresholds were low (median $r$ value = 0.40). Wysocki (personal communication) obtained reliability estimates for pyridine detection thresholds in an ascending method of limits procedure which ranged from $r = 0.44$ for sniff bottles ($n = 26$) to $r = 0.83$ for plastic squeeze bottles ($n = 29$). Although it would appear that the UPSIT is more reliable than such threshold measures, the reliability coefficients of this study cannot be directly compared with those noted above, since the latter are not based upon subjects deliberately selected to represent a wide range of initial test scores.

The finding of no statistically significant influence of release procedure (pencil tip versus sandpaper) upon the test scores is of practical importance, since some subjects,
when using sandpaper, eliminate the stimulus source by scratching the microencapsulated strips too vigorously. The use of a pencil tip obviates this problem. Furthermore, such use decreases the time necessary for the subject to complete the test, since the same instrument is employed to release the odorants and record the responses. These data suggest that clinicians and researchers using this instrument can now release the microencapsulated stimuli by pencil tip without significantly influencing the test scores.

Acknowledgements

We thank R.Gregg Settle for his comments on a previous version of the manuscript. This study was supported by National Institute of Neurological and Communicative Disorders and Stroke Grant PO 1 NS 16365.

References


Received September 1984; accepted December 1984