Perception & Psychophysics
1989, 45 (4), 381-384

Internal consistency reliability of the fractionated and whole University of Pennsylvania Smell Identification Test

RICHARD L. DOTY, RICHARD E. FRYE, and UDAYAN AGRAWAL
University of Pennsylvania, Philadelphia, Pennsylvania

The internal consistency reliability (ICR) of the 40-item University of Pennsylvania Smell Identification Test (UPSIT) and its 10-, 20-, and 30-item fractions was explored, as well as the relationships between the fractions and the entire 40-item test. Pearson correlation coefficients (*r*s) were computed among all independent combinations and permutations of the four 10-item UPSIT booklets using data from 774 subjects. The median *r* values of the 10- and 20-item combinations were used to establish the ICRs of the 10- and 20-item tests. The ICRs of the 30- and 40-item tests were estimated using the Spearman-Brown formula and the median *r*s of the 20-item combinations. Additional ICR estimates of the 40-item UPSIT were obtained from nonsymmetrical fractions using the Horst formula. The ICRs for the UPSIT and its 10-, 20-, and 30-item fractions were 0.922, 0.752, 0.855 and 0.898, respectively. No major sex differences emerged. Estimates of correlations between (1) single booklets and two-booklet combinations and (2) the 40-item UPSIT using Guilford's (1953) correction for nonindependence ranged from 0.812 to 0.871. Overall, these results indicate that (1) the UPSIT and its 10-, 20-, and 30-item fragments have very high ICRs and (2) individual UPSIT booklets or their combinations can be used to assess smell function in a reliable manner where extreme time constraints are present (e.g., in surveys and in brief neuropsychological test batteries).

Olfaction, compared to the other senses, has received little scientific attention from sensory psychologists and physiologists, largely as a consequence of the impracticality and time-consuming nature of traditional tests of smell function. Since no physical stimulus dimensions have been identified for odor quality analogous to those of wavelength for color or frequency for pitch, quantitative smell testing has focused upon manipulations of odorant concentration (e.g., detection threshold and suprathreshold scaling procedures). Because of the large number of chemicals available for testing, the requirement for precise stimulus control, and the lack of consensus as to the best psychophysical paradigm to employ, such measures have not been standardized across clinics or laboratories.

An alternative approach for the assessment of smell function derives from test measurement theory and focuses on the comparative ability of individuals to identify a number of odorants at the suprathreshold level. A test we developed using this approach—the University of Pennsylvania Smell Identification Test (UPSIT)—has made it possible to conveniently and accurately measure smell function in nonlaboratory settings without the use of complex olfactometric equipment or cumbersome sniff bottles (Doty, Shaman, Applebaum, et al., 1984; Doty, Shaman, & Dann, 1984). This 40-item "scratch & sniff" microencapsulated odorant test (commercially known as the Smell Identification Test®, Sensonics, Inc., Haddonfield, NJ) has gained wide acceptance within the medical and psychological communities and is now routinely administered in nearly 1,500 clinics in North America. In addition to serving as a validation criterion for other olfactory tests (e.g., Wright, 1987), the UPSIT has been shown to be sensitive to a wide range of smell deficits, including those due to: sinusitis and polyposis (e.g., B. W. Jafek, Moran, Eller, Rowley, & T. B. Jafek, 1987), industrial chemical exposure (Schwartz, Doty, Frye, Monroe, & Barker, in press), cystic fibrosis (Weiffenbach & McCarthy, 1984), Korsakoff's psychosis (Mair et al., 1986), Alzheimer's disease (Doty, Reyes, & Gregor, 1987; Warner, Peabody, Flattery, & Tinklenburg, 1986), parkinsonism (Doty, Deems, & Stellar, 1988), Kallman's syndrome (Doty, Shaman, & Dann, 1984), and lesions of the cerebral cortex (Jones-Gotman & Zatorre, 1988). Because this test can be self-administered and sent to subjects or patients through the mail, it is amenable to applications outside the traditional clinical or laboratory setting and allows for convenient longitudinal tracking of smell function.

An earlier study of 69 subjects suggested that the UPSIT has a high degree of odd/even item internal consistency reliability (ICR; $r = 0.93$) (Doty, Newhouse, & Azzalina, 1985). This work did not explore the relationships

This work was supported by Grant NS 16365 from the National Institute of Neurological and Communicative Disorders and Stroke. We are grateful for the editorial assistance of Carolyn Krasnow, and for the helpful suggestions of Daniel Deems, Judy Risser, and R. Gregg Settle. Correspondence concerning this article should be addressed to Richard L. Doty, Director, Smell and Taste Center, Hospital of the University of Pennsylvania, 3400 Spruce Street, Philadelphia, PA 19104.

among the different booklets of the UPSIT and did not determine whether subsections of the test can be used to assess smell function in cases where time constraints preclude the administration of the entire test battery (which requires 10–15 min). Situations where such a need exists include surveys, brief neuropsychological test batteries, and unusual experimental protocols. An example of the latter is a recent study of the influences of gravity on the olfactory function of a small group of subjects tilted in space (Mester, Doty, Shapiro, & Frye, 1988). Because many subjects felt uncomfortable being in the upside down position for more than a few minutes, it was not possible to administer the entire UPSIT on a single test occasion.

In the present study, we established Pearson correlations among all independent combinations and permutations of the four UPSIT test booklets, using data from a relatively large number of subjects. This allowed us to (1) ascertain the ICR of various subcomponents of the test, (2) more accurately determine the overall ICR of the UPSIT, and (3) establish correlations between each of the four 10-item booklets (and their combinations) and the entire test.

METHOD

Data Set

UPSIT data from 378 men and 396 women [respective mean (±*SD*) ages = 42.32 (25.40) and 50.36 (27.86) years] were randomly selected from the Smell and Taste Center's computerized data base with three constraints: (1) that they be from healthy individuals; (2) that roughly the same number of males and females be sampled; and (3) that the final data set represent approximately equal numbers of scores within the 7- to 40-item UPSIT range. Scores below 7 were not included, since they have a high probability of being from persons who are malingering (Doty, 1989; Doty, Shaman, Applebaum, et al., 1984).

Description of the UPSIT

Details of the University of Pennsylvania Smell Identification Test are presented elsewhere (Doty, 1989; Doty, Shaman, Applebaum, et al., 1984; Doty, Shaman, & Dann, 1984). Briefly, this test consists of four envelope-sized booklets, each containing 10 "scratch & sniff" odorants. The odorants are embedded in 10- to 50-μm urea-formaldehyde polymer microcapsules fixed in a proprietary binder and positioned on brown strips at the bottom of the pages of the test booklets. The stimuli are released by the scratching of the strips with a pencil tip in a standarized manner. Above each odorant strip is a multiple-choice question with four alternative responses for each item. For example, one of the items reads: "This odor smells most like: a) chocolate; b) banana; c) onion; or d) fruit punch." The test is forced-choice—that is, the subject is required to mark one of the four alternatives even if no smell is perceived. The criteria for the selection of the odorants and response alternatives are described elsewhere, as are the age- and gender-related norms, which are based upon several thousand subjects (Doty, 1989; Doty, Shaman, Applebaum, et al., 1984).

General Procedures

The number of correct responses for each UPSIT booklet was first calculated for all subjects. Pearson correlations were then computed among all independent combinations and permutations of the four 10-item UPSIT test booklets. For the 10- and 20-item UPSIT fractions, the ICR was estimated from the median *r* values of the independent 10- and 20-item fraction combinations. For the 30-item UPSIT fraction and for the total UPSIT, the ICR was estimated using the Spearman-Brown formula on the median of the 20-item split-half combinations (Guilford, 1954). This formula is

$$r_n = \frac{nr}{1 + (n-1)r},$$

where *r* is the original correlation and r_n = the reliability of the test *n* times as long.

In addition to estimating the whole-test ICR from the split-half correlations and the Spearman-Brown formula, we also estimated this reliability by using the single test booklets and their combinations and applying Horst's (1951) formula for unequal parts. This formula is

$$R = \frac{r[\sqrt{r^2 + 4pq(1-r^2)} - r]}{2pq(1-r^2)},$$

where *R* is the reliability of the whole test, *r* is the correlation between the two parts, *p* is the proportion of the total test devoted to one part, and *q* is $1-p$.

Since it is of interest to ascertain the relationship between the individual booklet score and the entire test score, we established the correlation between the scores of (1) each test booklet and the independent 2-booklet combinations and (2) the 40-item UPSIT, using Guilford's (1953) equation, which corrects for the spurious correlation resulting from lack of independence of these measures:

$$r_{ir} = \frac{r_{it}SD_t - SD_i}{\sqrt{(SD_i^2 + SD_t^2 - 2r_{it}SD_iSD_t)}}.$$

In this formula, r_{ir} is the corrected correlation, r_{it} is the correlation between the booklet and the total test score, SD_i is the standard deviation of the test booklet score, and SD_t is the standard deviation of the total test score. The correlation between the three-booklet combinations and the 40-item test could not be computed using this formula, since the formula overestimates the spurious component in cases where the fraction is larger than the remainder of the test. Clearly, however, the correlation of three booklets with the whole test would be greater than the correlation of the two test booklets with the test.

RESULTS

The Pearson correlation coefficients (*r*s) among the scores of the four UPSIT test booklets for both sexes are presented in Table 1 (all *p*s < .001). Analogous *r*s between the various independent combinations and permutations of the test booklets are presented in Table 2 (all *p*s < .001). It is apparent from these tables that (1) the *r* values are very similar for men and women, (2) relatively strong *r* values are present for the individual test booklets and their combinations, and (3) the *r* values increase as the length of the test fractions increase.

The median ICR values for the one-, two-, and three-booklet fractions are 0.752, 0.855, and 0.890, respectively. The median ICR of the whole UPSIT = 0.922. The ICRs for the entire UPSIT calculated using all possible independent booklet combinations and the Spearman-Brown and Horst (1951) formulae are shown in Table 3.

The estimates of the correlations of the one- and two-booklet test combinations with the whole UPSIT are presented in Table 4. It is apparent from this table that these fractions correlate highly with the overall UPSIT score (median *r* values for 10-item and 20-item tests = 0.818 and 0.855, respectively).

DISCUSSION

The present study demonstrates that the 40-item University of Pennsylvania Smell Identification Test and its 10-, 20-, and 30-item fractions have a high degree of internal consistency reliability. Furthermore, the data indicate that the aforementioned fractions correlate strongly with the overall UPSIT score, implying that they can be used, in appropriate instances, independently of the 40-item test. However, it should be emphasized that the 40-item test was designed to completely separate, for all practical purposes, the distributions of scores from anosmics and normal subjects and to provide a statistical basis for detecting malingering. Tests of lesser length will not achieve such a high degree of resolution, and care must be taken in any given case to determine whether the use of an UPSIT fraction is appropriate. Obviously, very high resolution is not as critical in protocols where only statistical differences among groups are sought. In clinical or industrial applications where smell dysfunction is suspected on the basis of an UPSIT fraction score, the complete 40-item battery can be subsequently administered to suspect cases to more accurately define the deficit. Because slight variations in average 10-item test scores are present among booklets, it would be prudent for workers using UPSIT fractions in experimental applications to either (1) counterbalance the use of such booklets or (2) use only one (or a given set) of the four booklets for within-study comparisons.

Table 1
Pearson Correlation Coefficients Among the Scores of the Four 10-item Test Booklets of the University of Pennsylvania Smell Identification Test (UPSIT)

	Booklet			
Booklet	1	2	3	4
		Total Group		
1	—	0.754	0.735	0.749
2		—	0.729	0.769
3			—	0.764
4				—
		Males Only		
1	—	0.750	0.725	0.758
2		—	0.729	0.777
3			—	0.771
4				—
		Females Only		
1	—	0.759	0.747	0.743
2		—	0.729	0.762
3			—	0.761
4				—

Table 2
Pearson Correlation Coefficients Among Scores on the Independent Combinations of the Four UPSIT Booklets

Booklet Combinations	Total Group	Males	Females
1&2 vs. 3&4	0.847	0.848	0.847
1&3 vs. 2&4	0.855	0.859	0.853
1&4 vs. 2&3	0.869	0.866	0.872
1 vs. 2,3,4	0.816	0.812	0.820
2 vs. 1,3,4	0.822	0.823	0.822
3 vs. 1,2,4	0.812	0.809	0.816
4 vs. 1,2,3	0.837	0.847	0.829

Table 3
Internal Consistency Reliabilities (ICRs) Estimated from Independent Combinations of the Four UPSIT Test Booklets

Booklet	ICR		
Combinations	Total Group	Males	Females
1&2 vs. 3&4	0.917	0.918	0.917
1&3 vs. 2&4	0.922	0.924	0.921
1&4 vs. 2&3	0.930	0.928	0.932
1 vs. 2,3,4	0.920	0.918	0.922
2 vs. 1,3,4	0.923	0.924	0.933
3 vs. 1,2,4	0.918	0.917	0.920
4 vs. 1,2,3	0.931	0.935	0.927
Median	0.922	0.924	0.922

Note—The Spearman-Brown formula was used to calculate the first three values above ($n = 2.0$); the Horst formula was used to calculate the next four ($p = .25$). See text for details.

Table 4
Correlations Between the Whole UPSIT and Individual Booklets and Two-Booklet Combinations

Books	Total Group	Males	Females
	Individual Booklets		
1	0.815	0.813	0.820
2	0.821	0.823	0.820
3	0.812	0.809	0.816
4	0.838	0.848	0.830
Median	0.818	0.818	0.820
	Two-Booklet Combinations		
1 & 2	0.845	0.849	0.844
2 & 3	0.870	0.866	0.873
3 & 4	0.850	0.849	0.850
1 & 4	0.871	0.867	0.871
2 & 4	0.856	0.858	0.853
1 & 3	0.854	0.860	0.853
Median	0.855	0.859	0.853

Note—Guilford's (1953) formula was used to correct for the spurious portion of the correlation due to lack of independence. See text for details.

The results of the present study compare favorably with those of the earlier study of 69 subjects in which an odd-even paradigm was used to calculate ICR for the entire UPSIT (Doty et al., 1985). In the present case, the estimated 40-item UPSIT ICR was 0.922, whereas in the earlier work this value was 0.930. Thus, it appears that the 40-item UPSIT ICR is slightly above 0.92.

The Spearman-Brown formula appears to be an accurate predictor of UPSIT ICRs, as indicated by a comparison of our empirically determined ICRs with ones predicated on the assumptions of this formula. Thus, we found the median empirical correlation among the 20-item fractions of the UPSIT to be 0.855. Given the median correlation among the 10-item fractions ($r = 0.752$) and the Spearman-Brown formula, the predicted value for the correlation among the 20-item fractions is 0.858, essentially the same as that which was established empirically.

The finding that the fractions of the UPSIT have strong internal consistency reliabilities and that the sectors of the whole test correlate well with one another suggest that the test items provide a measure of a rather homogeneous trait. This observation, along with evidence that relatively high correlations are present among clinical detection thresholds for a number of odorants, lends support to Yoshida's (1984) notion that a "general olfactory acuity" factor exists which is analogous to the general intelligence factor derived from items of intelligence tests. Such a conceptualization does not, however, preclude the possibility of the presence of specific olfactory factors, such as those which may be associated with specific anosmias or hyposmias (see, e.g., Amoore, 1971).

REFERENCES

Amoore, J. E. (1971). Olfactory genetics and anosmia. In L. M. Beidler (Ed.), *Handbook of sensory physiology: Vol. 4. Chemical Senses: Part 1. Olfaction* (pp. 245-256). New York: Springer-Verlag.

Doty, R. L. (1989). *The Smell Identification Test™ administration manual* (2nd ed.). Haddonfield, NJ: Sensonics, Inc.

Doty, R. L., Deems, D. A., & Stellar, S. (1988). Olfactory dysfunction in Parkinson's disease: A general deficit unrelated to neurologic signs, disease stage, or disease duration. *Neurology*, **38**, 1237-1244.

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

Doty, R. L., Reyes, P. F., & Gregor, T. (1987). Presence of both odor identification and detection deficits in Alzheimer's disease. *Brain Research Bulletin*, **18**, 597-600.

Doty, R. L., Shaman, P., Applebaum, S. L., Giberson, R., Sikorsky, L., & Rosenberg, L. (1984). Smell identification ability: Changes with age. *Science*, **226**, 1441-1443.

Doty, R. L., Shaman, P., & Dann, M. (1984). Development of the University of Pennsylvania Smell Identification Test: A standardized microencapsulated test of olfactory function. *Physiology & Behavior (Monograph)*, **32**, 489-502.

Guilford, J. P. (1953). The correlation of an item with a composite of the remaining items in a test. *Educational & Psychological Measurement*, **13**, 87-93.

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

Horst, P. (1951). Estimating total test reliability from parts of unequal length. *Educational & Psychological Measurement*, **11**, 368-371.

Jafek, B. W., Moran, D. T., Eller, P. M., Rowley, J. C., & Jafek, T. B. (1987). Steroid-dependent anosmia. *Archives of Otolaryngology–Head & Neck Surgery*, **113**, 547-549.

Jones-Gotman, M., & Zatorre, R. J. (1988). Olfactory identification deficits in patients with focal cerebral excision. *Neuropsychologia*, **26**, 387-400.

Mair, R. G., Doty, R. L., Kelly, K. M., Wilson, C. S., Langlais, P. J., McEntee, W. J., & Vollmecke, T. A. (1986). Multimodal sensory discrimination deficits in Korsakoff's psychosis. *Neuropsychologia*, **24**, 831-839.

Mester, A. F., Doty, R. L., Shapiro, A., & Frye, R. E. (1988). Influence of body tilt within the sagittal plane on olfactory function. *Aviation, Space, & Environmental Medicine*, **59**, 734-737.

Schwartz, B., Doty, R. L., Monroe, C., Frye, R. E., & Barker, S. (in press). Olfactory function in chemical workers exposed to acrylate and methacrylate vapors. *American Journal of Public Health*.

Warner, M. D., Peabody, C. A., Flattery, J. J., & Tinklenberg, J. R. (1986). Olfactory deficits and Alzheimer's disease. *Biological Psychiatry*, **21**, 116-118.

Weiffenbach, J. M., McCarthy, V. P. (1984). Olfactory deficits in cystic fibrosis: Distribution and severity. *Chemical Senses*, **9**, 193-199.

Wright, H. N. (1987). Characterization of olfactory dysfunction. *Archives of Otolaryngology–Head & Neck Surgery*, **113**, 163-168.

Yoshida, M. (1984). Correlation analysis of detection threshold data for "standard test" odors. *Bulletin of the Faculty of Science & Engineering of Chuo University*, **27**, 343-353.

(Manuscript received August 4, 1988;
revision accepted for publication October 12, 1988.)