Chemical Senses, 2019, Vol 44, 365–369
doi:10.1093/chemse/bjz025
Original Article
Advance Access publication 9 May 2019

Original Article

Long-Term Test–Retest Reliability of the UPSIT in Cognitively Intact Older Adults

Davangere P. Devanand[1,2], Xinhua Liu[3], Hannah Cohen[1], John Budrow[1,4], Nicole Schupf[5], Jennifer Manly[5] and Seonjoo Lee[2,3,6]

[1]Division of Geriatric Psychiatry, New York State Psychiatric Institute, New York, NY, USA, [2]Department of Psychiatry, Columbia University Medical Center, New York, NY, USA, [3]Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY, USA, [4]New York Medical College, Valhalla, NY, USA, [5]Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University, New York, NY, USA and [6]Division of Mental Health Data Science, New York State Psychiatric Institute, New York, NY, USA

Correspondence to be sent to: Davangere P. Devanand, Division of Geriatric Psychiatry, New York State Psychiatric Institute, 1051 Riverside Drive, Unit 126, New York, NY 10032, USA. e-mail: dpd3@cumc.columbia.edu

Editorial Decision 27 April 2019.

Abstract

The objective of this study was to determine the long-term test–retest reliability of the University of Pennsylvania Smell Identification Test (UPSIT), and its individual items, in cognitively intact older adults. A community sample of older adults received a neuropsychological test battery, including the 12-item, 6-trial Selective Reminding Test (SRT). The UPSIT was administered at baseline and follow-up that occurred between 1 and 4 years after baseline. UPSIT scores of participants who were cognitively intact and did not decline cognitively were examined for test–retest reliability. In 92 older adults with mean age 77.6 years followed for 2.79 (standard deviation [SD] 0.69) years, mean UPSIT score declined from 30.29 (SD 5.83) to 27.80 (SD 5.50). In linear mixed models that adjusted for time, age, sex, and education, intraclass correlation coefficients for UPSIT were 0.65, SRT delayed recall 0.59, and SRT total immediate recall 0.49. Among 4 possible response combinations, the largest proportion of participants had correct responses at both visits for 35 out of 40 items. Consistency of item responses ranged from 50% to 90% across the 2 time points. The long-term test–retest reliability of the UPSIT was moderately strong without practice effects over long periods of time in older adults. These results provide indirect support to prior findings on odor identification impairment predicting cognitive decline and dementia, and suggest potential use of olfactory testing as a biomarker in prevention and treatment trials of cognitive enhancers.

Key words: aging, item analysis, odor identification, olfaction, reliability, test–retest

Introduction

The olfactory bulb, and olfactory projection pathways to limbic brain regions, are infiltrated by neurofibrillary tangles and, to a lesser extent, amyloid plaques, both of which are the pathologic hallmarks of Alzheimer's disease (AD) (Hyman et al. 1991). Impaired odor identification is a well-replicated marker of olfactory deficits in patients with AD. Of 30 published studies, all showed impaired odor identification in AD compared with healthy control subjects (Sun et al. 2012). These deficits also predict the transition from mild cognitive impairment to AD and an increased likelihood of cognitive decline in cognitively intact middle-aged and older individuals (Stanciu et al. 2014; Devanand et al. 2015).

The University of Pennsylvania Smell Identification Test (UPSIT) is a widely used assessment of olfactory ability that has been reported to have a high test–retest reliability at a 2-week interval (r = 0.95) and at a 6-month interval (r = 0.92) in young to middle-aged adults

with intact cognition (Doty, Shaman, Dann 1984; Doty et al. 1985). Short-term test–retest reliability for the UPSIT has been shown to compare favorably with several other odor identification tests (Doty et al. 1995). For the Sniffin' Sticks test, which is a widely used olfaction test developed in Europe, test–retest reliability over a 1-day interval was moderate with Pearson's $r = 0.79$ for the 16-odor identification version (Schriever et al. 2011) and ranged from moderate to strong (Pearson's $r = 0.69$–0.93) for short and long versions of that test (Haehner et al. 2009).

There is an absence of information on the long-term test–retest reliability of the UPSIT in cognitively intact older adults where the test has been shown to predict cognitive decline and incident AD during long-term follow-up (Devanand et al. 2008, 2015). Furthermore, the potential utility of odor identification testing for diagnosis and estimating prognosis in cognitively impaired older adults may be confounded by the fact that odor identification ability worsens with aging, particularly in the later decades of life (Doty, Shaman, Applebaum, et al. 1984). Therefore, it is important to ascertain if there is high test–retest reliability for the UPSIT over an extended period of time in cognitively intact older adults. The objective of this study was to evaluate the long-term test–retest reliability for the UPSIT total score in cognitively intact older individuals participating in a multiethnic older community cohort study with long-term follow-up (Devanand et al. 2015). We also explored the test–retest reliability of the 12-item B-SIT, a widely used subscale of the UPSIT, and the 40 specific odors that comprise the UPSIT.

Materials and methods

The Washington Heights/Inwood Columbia Aging Project (WHICAP) consists of individuals recruited from a stratified random sample of 50% of all Medicare beneficiaries aged 65 years and older in northern Manhattan, New York City. Participants were recruited in 1992 (approximately 25% of participants) as well as between 1999 and 2001 (approximately 75% of participants). The Columbia University Institutional Review Board approved the study protocol and written informed consent forms. The study complies with the Declaration of Helsinki for medical research involving human subjects.

All participants received a standardized neuropsychological test battery at each visit that included measures of learning and memory, orientation, abstract reasoning, executive function, language, and visuospatial ability. Odor identification testing was performed with the UPSIT. The UPSIT comprises 40 common odorants embedded in microcapsules with 1 odor on each page. On each page, the participant scratches the odorant strip, smells the odor emitted from the microcapsule, and then chooses the best answer among 4 multiple choice items for each emitted odor. The total UPSIT score ranges from 10 to 40. Participants were tested in English or Spanish based on their stated language preference. The Spanish version of the UPSIT differs from the English version by 5 odors. The remaining 35 odors are the same odors as in the English UPSIT, and all items use Spanish word labels. The UPSIT was first administered at the evaluation between 2004 and 2006 with follow-up UPSIT testing that took place 1–4 years after the initial UPSIT administration.

For inclusion in this study, participants needed to complete the UPSIT at baseline and follow-up and to be classified as not having dementia at both of these time points. Furthermore, to avoid the confound of cognitive decline being associated with lower UPSIT scores over time, participants needed to have intact cognitive ability at baseline and lack of cognitive decline during follow-up. Intact cognitive ability at baseline was defined a priori as a score ≤ 2 on a test of global cognition, the 28-item Blessed Orientation Memory Concentration Test (brief test of global cognition, range 0–28, higher score indicates worse cognition) together with a baseline 12-item, 6-trial Selective Reminding Test (SRT) delayed recall score ≥ 7 out of 12 (range 0–12, higher score indicates better recall; Buschke 1973). The SRT measures verbal learning and memory through the use of a list-learning procedure of 12 words over 6 trials. Lack of cognitive decline was defined as change in SRT delayed recall score over time being less than half the square root of the mean square error of the decline estimated by the model for the change.

Statistical Analyses

There were 283 study participants who met criteria for intact cognition at baseline, were not diagnosed with dementia at follow-up, and had SRT and UPSIT test scores available at baseline and follow-up. A linear regression model examined the effect of baseline age, sex, education, follow-up time, and baseline SRT delayed recall test score on the within-person change of the test score over time. Using the residual of the fitted model, covariate-adjusted change in the SRT delayed recall test score was estimated for each individual. Subjects with estimated decline more than half the square root of the mean square error were considered as having cognitive decline and were excluded, resulting in a final study sample of 92 study participants.

To describe the demographic and clinical characteristics of the study sample, we calculated summary statistics with mean and SD for continuous variables and percent for categorical variables that included the response pattern of individual UPSIT items at the 2 time points. To evaluate reliability of memory and olfactory tests at baseline and follow-up, we calculated the intraclass correlation coefficient (ICC) using linear mixed effect models, with and without the fixed effect of covariates of time of follow-up, baseline age, sex, education, or language of test administration.

The responses of UPSIT items at 2 time points have 4 possible categories, with (0, 1) for incorrect at baseline but correct at follow-up, (1, 0) for correct at baseline but incorrect at follow-up, (0, 0) for incorrect at both visits, and (1, 1) for correct responses at both visits. As each odor has 4 choices with one correct answer, the likelihood of choosing a correct answer by chance is 25%. Thus, the item response (1, 1) is likely to reveal a stable ability of odor recognition, whereas the other response categories might be affected by random error. Furthermore, the item responses (0, 0) could be the result of an irreversible deficit of odor identification or from a random choice at the time. The item response of (1, 0) could be due to decline in the ability of odor recognition or possible random error, and the item response of (0, 1) may imply an unstable ability of odor recognition or a random choice. To identify the items of the UPSIT associated with stable odor recognition, we applied a logistic model to examine the association between the binary outcome of correct response at both visits, specifically (1, 1) versus the other 3 possibilities for each item, and follow-up time, controlling for baseline, age, and sex.

Results

Demographic and clinical characteristics of the study sample of 92 older adults are described in Table 1. English speakers tested with the English UPSIT comprised 81.5% and Spanish speakers tested with the Spanish UPSIT comprised 18.5% of the sample. The majority were female (59.6%) and the baseline age range was 70–91 years with mean 77.55 (standard deviation [SD] 4.49) years. Follow-up ranged from 1.19 to 3.99 years with mean follow-up time of 2.79

(SD 0.69) years. Mean UPSIT score was 30.29 (SD 5.83) at baseline and 27.80 (SD 5.50) at follow-up, with a mean decrease of 2.48 (SD 4.65). Partly as a result of the selection criteria requiring good performance on the SRT, the mean declines in episodic verbal memory during follow-up were small: 0.04 (SD = 1.35) for SRT delayed recall and 0.64 (SD = 6.52) for SRT total immediate recall. UPSIT score was positively correlated with SRT total immediate recall at baseline (r = 0.27, P = 0.0082) but the correlation attenuated at follow-up (r = 0.17, P = 0.10). In contrast, UPSIT score did not show significant associations with SRT delayed recall at either baseline (r = 0.16, P = 0.13) or follow-up (r = 0.19, P = 0.075) (Table 1).

Test–retest reliability of the total UPSIT score

The ICCs for the UPSIT and SRT measures were estimated based on linear mixed models, with and without adjusting for covariates, using 184 observations from 92 subjects. The estimated ICCs and coefficients of covariates for fixed effects are listed in Table 2. UPSIT scores significantly decreased with follow-up time and baseline age, were lower in males than females, and were unrelated to education or language. SRT total immediate recall and SRT delayed recall scores both did not change with follow-up time and were lower in older subjects and males. Spanish speakers had less education in years than English speakers. When both language and education were included in the models for the 2 SRT tests, education became statistically insignificant.

Without covariate adjustment, the total UPSIT score had a crude ICC of 0.59, similar to ICC of 0.58 for SRT immediate recall, and both were lower than the ICC of 0.67 for SRT delayed recall. When adjusting for follow-up time only, the ICC for the UPSIT was 0.67, similar to the ICC of 0.67 for SRT delayed recall, and higher than the ICC of 0.59 for SRT total immediate recall. In linear mixed effects models that adjusted for specific demographic covariates as described in Table 2, the ICC for the UPSIT was 0.65 in all 3 models, the ICC for SRT delayed recall showed reduced values between 0.58 and 0.61, and the ICC for SRT total immediate recall showed even lower values between 0.47 and 0.52 (Table 2).

Table 1. Demographic and clinical characteristics of the sample (n = 92)

Variable	Mean (SD)	Range
Age at baseline (years)	77.55 (4.49)	70–91
Years of follow-up	2.79 (0.69)	1.19–3.99
Total UPSIT score at baseline	30.29 (5.83)	11–40
Total UPSIT score at follow-up	27.80 (5.50)	13–37
SRT total recall baseline	49.35 (7.00)	26–65
SRT delayed recall baseline	8.52 (1.35)	7–12
Education in years	13.62 (3.65)	3–20
Test administered in Spanish %	18.48	—
Male %	30.43	—
White %	57.61	—
African American %	18.48	
Hispanic %	23.91	

Test–retest reliability of UPSIT items

The standardized Cronbach alpha coefficient for internal consistency of UPSIT items was 0.83 at baseline and 0.78 at follow-up. In the minority of participants who were tested in Spanish, their UPSIT scores at baseline and change over time did not differ significantly from the English version of the UPSIT.

For item-specific responses, for the 4 pairs of response combinations, the largest proportion of participants had correct responses at both visits (1, 1) for 35 out of 40 items, with 26 items having the proportion of (1, 1) ranging between 53.26% and 89.13%. Consistency of responses, defined as either 0–0 (incorrect–incorrect) or 1–1 (correct–correct), ranged from 50% to 90.22% across all items, that is, all items had more consistent than inconsistent (0–1 or 1–0) responses across the 2 time points. Three items (dill pickle, lime, and grass) had their proportion of correct responses at both time points (1, 1) ranging from 29.35% to 32.61%, similar to the proportion of responses for both incorrect (0, 0) with values ranging from 29.35% to 35.87%. Only 2 items had a low proportion of correct responses at both time points (1, 1), 18.48% and 17.39% for Cheddar Cheese and Lemon, respectively, which were lower than 31.52% and 43.48% for incorrect responses (0, 0) for these 2 items. Item-specific logistic regression analysis indicated that age- and sex-adjusted odds of correct responses (1, 1) declined significantly with follow-up time for 3 items (cherry [P = 0.018], leather [P = 0.072], and lemon [P =0.074]), whereas it improved for wintergreen (P = 0.021). The odds of correct responses were unrelated to

Table 2. ICC for test–retest reliability, based on linear mixed effect models

	SRT total immediate recall		SRT delayed recall		UPSIT	
Covariates	*B* (SE)	ICC	*B* (SE)	ICC	*B* (SE)	ICC
None	—	0.58	—	0.67	—	0.59
Time	−0.37 (0.23)	0.59	−0.06 (0.05)	0.67	−0.90 (0.17)	0.67
Time	−0.36 (0.23)	0.52	−0.05 (0.05)	0.61	−0.89 (0.17)****	0.65
Baseline age	−0.54 (0.14)***		−0.12 (0.03)***		−0.23 (0.12)$^{+}$	
Male vs. female	−4.30 (1.33)**		−1.10 (.03)***		−2.59 (1.15)*	
Time	−0.35 (0.23)	0.49	−0.05 (0.05)	0.59	−0.89 (0.17)****	0.65
Baseline age	−0.46 (0.13)**		−0.10 (0.03)**		−0.22 (0.12)$^{+}$	
Male vs. female	−4.77 (1.29)***		−1.19 (0.31)***		−2.66 (1.16)*	
Education	0.47 (0.16)**		0.09 (0.04)*		0.07 (0.15)	
Time	−0.35 (0.23)	0.47	−0.05 (0.05)	0.58	−0.89 (0.17)****	0.65
Baseline age	−0.44 (0.13)**		−0.10 (0.03)**		−0.23 (0.12)$^{+}$	
Male vs. female	−4.12 (1.26)**		−1.07 (0.30)***		−2.58 (1.15)*	
Spanish vs. English	−4.97 (1.49)**		−0.95 (0.36)**		−0.33 (1.36)	

B, estimated coefficient of covariate; SE, standard error.

$^{+}0.05 < P < 0.07$, $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$.

follow-up time (P's ≥ 0.098) for the remaining 36 items. The results indicated that instability of correct responses was present for only 4 out of 40 UPSIT items.

For the B-SIT, ICCs (crude ICC: 0.42, covariate-adjusted ICC: 0.45–0.49) were not greater than the ICCs (crude ICC: 0.60, covariate-adjusted ICC: 0.64–0.66) of a subscale with the remaining 28 UPSIT items.

Discussion

In a community sample of older individuals who were cognitively intact and did not decline cognitively, UPSIT scores obtained an average of 2.79 years apart (range 1–4 years) showed moderately strong test–retest reliability that was lower than has been reported for shorter intervals of 2 weeks to 6 months. This is consistent with the expectation of decreased reliability with an increasing time interval between test administrations (Doty, Shaman, Dann 1984; Doty et al. 1985). The decline in UPSIT scores during follow-up indicates lack of a practice effect, which is consistent with lack of a practice effect at shorter time intervals of 2 weeks and 6 months between test administrations (Doty, Shaman, Dann 1984; Doty et al. 1985). UPSIT test–retest reliability improved after controlling for age, sex, and time between administrations of the UPSIT. These findings support the utility of the UPSIT as a reliable test not only over short time intervals as previously demonstrated, but also over long intervals in cognitively intact older adults, thereby providing indirect support for potential clinical and research utility.

When the UPSIT was developed, internal consistency reliability (ICR) was estimated from the median of bivariate correlations of 10- and 20-item scores for combinations of booklets based on cross-sectional data. ICR was 0.93 for the total UPSIT score and ranged from 0.73 to 0.90 for the 4 individual booklets and combinations of booklets (Doty et al. 1985). In our sample of cognitively intact older adults who did not decline cognitively, internal consistency was strong, and participants tested in Spanish had scores similar to participants tested in English (Devanand et al. 2010).

This sample averaged 78 years of age at baseline. The decline in UPSIT scores over time in this elderly sample is consistent with odor identification ability worsening with aging, particularly in the later decades of life (Doty, Shaman, Dann 1984). Other studies, including reports from the larger WHICAP cohort, show moderate positive correlations between odor identification and episodic verbal memory test scores (Devanand et al. 2010; Devanand et al. 2015; Woodward et al. 2018). In this WHICAP substudy, the correlations between the UPSIT and memory tests were relatively weak, which may be related to the inclusion criteria restricting the range of episodic verbal memory test scores (Devanand et al. 2010, 2015).

For every UPSIT item, the proportion of a consistent correct response at the 2 time points (correct–correct) exceeded 50% for 29 out of 40 items but was not close to 100% for any item. Of note, the 12-item B-SIT is a component of the 40-item UPSIT and is a shorter, more practical version with cross-cultural validation (Sun et al. 2012; Woodward et al. 2018). Nonetheless, test–retest reliability was not stronger for these 12 items in the B-SIT compared to the remaining 28 UPSIT items.

The normal physiology of olfactory pathways may explain the inability to distinguish specific UPSIT items with strong versus poor test–retest reliability. The experience of smelling an odor is the end result of small molecules that enter the nasal cavity, dissolve in the mucosa of the olfactory epithelium, and then interact with olfactory receptor neurons via transmembrane G-protein coupled olfactory receptor proteins. Most familiar odors contain several proteins and other molecules leading to considerable overlap and difficulty in discriminating between odors that are similar, for example, apple and orange (Doty 2017). The chemical complexity underlying each odor may not match well with the olfactory receptors and their projections in the olfactory pathways, leading to difficulty in replicating discriminative and predictive utility for individual odors even though overall odor identification ability can be reliably assessed with the 40-item UPSIT.

Odor identification testing can help to discriminate between diagnostic groups along the spectrum of cognitive decline and to predict cognitive decline and dementia broadly, as well as AD specifically (Devanand et al. 2015). Odor identification test performance is a biomarker that adds to the information obtained by clinical assessment and the use of other biomarkers for the early detection of AD (Devanand et al. 2008). It may also prove to be useful as a biomarker in selecting or stratifying patients in treatment trials of cognitively impaired individuals and in prevention trials because of its ability to predict cognitive decline in cognitively intact individuals (Devanand et al. 2015). These are important new approaches under investigation in the treatment of AD and its prodromal state. Assessment of odor identification, which is shown here to be reliable and with aging effects but without long-term practice effects in cognitively intact individuals, has potential utility in this process.

Funding

This work was supported by the following grants from the National Institute on Aging of the National Institutes of Health (P01AG07232, R01AG041795, R01AG057898, and R01AG058767).

Conflict of interest

D.P.D. is a consultant to Acadia, Avanir, Eisai, Genentech, and Neuronix and received research support from the NIH. X.L., H.C., and J.B. report no conflicts of interest. N.S. received research support from the NIH and Alzheimer's Association. J.M. received travel funding from the Alzheimer's Association; served as an associate editor for the *Journal of the International Neuropsychological Society*; consulted for National Academies of Medicine and National Academy of Sciences; and received research support from NIA, NIDDK, and NINDS. S.L. received research support from the NIH.

References

Buschke H. 1973. Selective reminding for analysis of memory and learning. *J Verbal Learning Verbal Behav*. 12(5):543–550.

Devanand DP, Lee S, Manly J, Andrews H, Schupf N, Doty RL, Stern Y, Zahodne LB, Louis ED, Mayeux R. 2015. Olfactory deficits predict cognitive decline and Alzheimer dementia in an urban community. *Neurology*. 84(2):182–189.

Devanand DP, Liu X, Tabert MH, Pradhaban G, Cuasay K, Bell K, de Leon MJ, Doty RL, Stern Y, Pelton GH. 2008. Combining early markers strongly predicts conversion from mild cognitive impairment to Alzheimer's disease. *Biol Psychiatry*. 64:871–879.

Devanand DP, Tabert MH, Cuasay K, Manly JJ, Schupf N, Brickman AM, Andrews H, Brown TR, DeCarli C, Mayeux R. 2010. Olfactory identification deficits and MCI in a multi-ethnic elderly community sample. *Neurobiol Aging*. 31:1593–1600.

Doty RL. 2017. Olfactory dysfunction in neurodegenerative diseases: is there a common pathological substrate? *Lancet Neurol*. 16:478–488.

Doty RL, McKeown DA, Lee WW, Shaman P. 1995. A study of the test-retest reliability of ten olfactory tests. *Chem Senses*. 20:645–656.

Doty RL, Newhouse MG, Azzalina JD. 1985. Internal consistency and short-term test-retest reliability of the University of Pennsylvania Smell Identification Test. *Chem Senses*. 10(3):297–300.

Doty RL, Shaman P, Applebaum SL, Giberson R, Siksorski L, Rosenberg L. 1984. Smell identification ability: changes with age. *Science*. 226:1441–1443.

Doty RL, Shaman P, Dann M. 1984. Development of the University of Pennsylvania Smell Identification Test: a standardized microencapsulated test of olfactory function. *Physiol Behav*. 32:489–502.

Haehner A, Mayer AM, Landis BN, Pournaras I, Lill K, Gudziol V, Hummel T. 2009. High test-retest reliability of the extended version of the "Sniffin' Sticks" test. *Chem Senses*. 34:705–711.

Hyman BT, Arriagada PV, Van Hoesen GW. 1991. Pathologic changes in the olfactory system in aging and Alzheimer's disease. *Ann N Y Acad Sci*. 640:14–19.

Schriever VA, Körner J, Beyer R, Viana S, Seo HS. 2011. A computer-controlled olfactometer for a self-administered odor identification test. *Eur Arch Otorhinolaryngol*. 268:1293–1297.

Stanciu I, Larsson M, Nordin S, Adolfsson R, Nilsson LG, Olofsson JK. 2014. Olfactory impairment and subjective olfactory complaints independently predict conversion to dementia: a longitudinal, population-based study. *J Int Neuropsychol Soc*. 20:209–217.

Sun GH, Raji CA, Maceachern MP, Burke JF. 2012. Olfactory identification testing as a predictor of the development of Alzheimer's dementia: a systematic review. *Laryngoscope*. 122:1455–1462.

Woodward MR, Hafeez MU, Qi Q, Riaz A, Benedict RHB, Yan L, Szigeti K; Texas Alzheimer's Research, Care Consortium. 2018. Odorant item specific olfactory identification deficit may differentiate Alzheimer disease from aging. *Am J Geriatr Psychiatry*. 26:835–846.