SMELL-S and SMELL-R: Olfactory tests not influenced by odor-specific insensitivity or prior olfactory experience

Julien W. Hsieh[a,b,1], Andreas Keller[a], Michele Wong[a], Rong-San Jiang[c], and Leslie B. Vosshall[a,d,e,1]

[a]Laboratory of Neurogenetics and Behavior, The Rockefeller University, New York, NY 10065; [b]Rhinology–Olfactology Unit, Service of Otorhinolaryngology Head and Neck Surgery, Department of Clinical Neurosciences, Geneva University Hospitals, CH-1211 Geneva 14, Switzerland; [c]Department of Otolaryngology, Taichung Veterans General Hospital, Xitun District, Taichung City, Taiwan 407; [d]Howard Hughes Medical Institute, New York, NY 10065; and [e]Kavli Neural Systems Institute, The Rockefeller University, New York, NY 10065

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2015.

Contributed by Leslie B. Vosshall, September 12, 2017 (sent for review June 26, 2017; reviewed by Avery Gilbert and Thomas Hummel)

Smell dysfunction is a common and underdiagnosed medical condition that can have serious consequences. It is also an early biomarker of neurodegenerative diseases, including Alzheimer's disease, where olfactory deficits precede detectable memory loss. Clinical tests that evaluate the sense of smell face two major challenges. First, human sensitivity to individual odorants varies significantly, so test results may be unreliable in people with low sensitivity to a test odorant but an otherwise normal sense of smell. Second, prior familiarity with odor stimuli can bias smell test performance. We have developed nonsemantic tests for olfactory sensitivity (SMELL-S) and olfactory resolution (SMELL-R) that use mixtures of odorants that have unfamiliar smells. The tests can be self-administered by healthy individuals with minimal training and show high test–retest reliability. Because SMELL-S uses odor mixtures rather than a single molecule, odor-specific insensitivity is averaged out, and the test accurately distinguished people with normal and dysfunctional smell. SMELL-R is a discrimination test in which the difference between two stimulus mixtures can be altered stepwise. This is an advance over current discrimination tests, which ask subjects to discriminate monomolecular odorants whose difference in odor cannot be quantified. SMELL-R showed significantly less bias in scores between North American and Taiwanese subjects than conventional semantically based smell tests that need to be adapted to different languages and cultures. Based on these proof-of-principle results in healthy individuals, we predict that SMELL-S and SMELL-R will be broadly effective in diagnosing smell dysfunction.

olfaction | smell test | hyposmia | olfactory dysfunction | odor-specific insensitivity

Smell dysfunction manifests itself primarily in the reduced ability to detect or distinguish volatile chemicals. It ranges from the complete inability to smell any odors to a partial reduction in olfactory sensitivity to smell distortion, for instance, a condition in which a large number of volatiles smell like cigarette smoke. The prevalence of smell dysfunction in the general adult population is about 20% in Europe and the United States (1–3). This condition is potentially dangerous because those affected are unable to detect fire, spoiled food, hazardous chemicals, and leaks of odorized natural gas (4, 5). Smell loss may give rise to health problems, including mental health symptoms such as depression, anxiety, and social isolation. It affects quality of life by altering food preferences and the amount of food ingested (5). Food is often perceived as bland or tasteless by patients with smell disorders, leading to loss of appetite or overeating (4, 5).

The most frequent causes of smell dysfunction are sinonasal diseases, upper respiratory tract infection, and head trauma. Smell loss can be congenital (6, 7), and in many cases, the cause is unknown (5, 8). Importantly, smell dysfunction is an early sign of Alzheimer's disease (9), the most common cause of dementia in the United States that is projected to affect an estimated 1 in every 45 individuals by 2050 (10). It is well established that diminished olfactory function arises early in the progression of Alzheimer's disease and is highly predictive of future cognitive decline (9, 11). Because of the high prevalence and dramatic consequences of smell loss, accurate diagnosis of olfactory dysfunction is important. While self-reported hearing loss tends to be accurate (12), self-reporting of olfactory dysfunction is notoriously unreliable (13, 14). Therefore, accurate diagnostic tests for smell dysfunction that can be deployed worldwide are critically important. Following a diagnosis, therapeutic options and counseling can be offered to patients suffering from smell loss (15).

In clinical smell testing, patients are presented with odor stimuli in a variety of formats, including scratch 'n' sniff strips, glass vials or jars, felt-tip pens, or paper scent strips used in perfume shops, and asked to answer questions about what they smell. Smell tests assess the ability of subjects to detect, discriminate, or identify odors. Olfactory threshold tests measure the lowest concentration of an odor stimulus that a patient can perceive, while discrimination tests assess the ability of subjects to distinguish two different smells. Finally, odor identification tests evaluate whether a patient can detect and match odors to standard words that describe the smell (16).

Significance

Currently available smell testing methods can be confounded by the lack of prior experience or insensitivity to the odorants used in the test. This introduces a source of bias into clinical tests aimed at detecting patients with olfactory dysfunction. We have developed smell tests that use mixtures of 30 molecules that average out the variability in sensitivity to individual molecules. Because these mixtures have an unfamiliar smell and the tests are nonsemantic, their use eliminates differences in test performance due to familiarity with the smells or the words used to describe them. SMELL-S and SMELL-R facilitate smell testing in different populations, without the need to adapt test stimuli to account for differences in familiarity with the test odors.

Author contributions: J.W.H., A.K., R.-S.J., and L.B.V. designed research; J.W.H. and M.W. performed research; J.W.H., A.K., and M.W. analyzed data; and J.W.H., A.K., and L.B.V. wrote the paper.

Reviewers: A.G., Synesthetics, Inc.; and T.H., Technical University of Dresden.

Conflict of interest statement: J.W.H., A.K., and L.B.V. are inventors on US provisional patent application 62/528,420, filed July 3, 2017, by The Rockefeller University, relating to the smell test methods in this manuscript.

[1]To whom correspondence may be addressed. Email: hsiehjulien@gmail.com or Leslie.Vosshall@rockefeller.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1711415114/-/DCSupplemental.

There are two major challenges to reliably testing a patient's sense of smell. First, sensitivity to monomolecular odorants varies greatly even among subjects with a normal sense of smell (17–19). If a patient has a low score on a test that assesses olfactory sensitivity with the rose-like odor phenylethyl alcohol (20), it is difficult to know whether the patient suffers from general smell dysfunction or is merely insensitive to phenylethyl alcohol with an otherwise normal sense of smell.

The second challenge is to develop a test that is not influenced by the patient's personal or cultural experiences with the test odorants. This has an obvious influence on the results of odor identification tests, such as the University of Pennsylvania Smell Identification Test (UPSIT), for which subjects are given a booklet with 40 scratch 'n' sniff items and asked to select one of four words (for example, "gingerbread," "menthol," "apple," or "cheddar cheese") that best describes what the odor smells like. Whether a patient can correctly identify the smell of gingerbread depends not only on the patient's sense of smell but also on whether the patient has previously encountered the smell of gingerbread. This in turn depends on many factors, such as the cultural and age group to which the person belongs. To address this familiarity problem, the UPSIT has been adapted for use in a number of countries worldwide by replacing unfamiliar items and adapting the answers on the multiple-choice test. For instance, the North American UPSIT was adapted for Taiwanese subjects by replacing "clove," "cheddar cheese," "cinnamon," "gingerbread," "dill pickle," "lime," "wintergreen," and "grass" with "sandalwood," "fish," "coffee," "rubber tire," "jasmine," "grapefruit," "magnolia," and "baby powder" (21). The strong influence of culture on such test results limits the utility of odor identification tests. Even performance on nonsemantic odor discrimination tasks depends on prior experience with the odorants (22, 23), and it is therefore important to avoid stimuli having differential familiarity in the test population.

We have developed two nonsemantic smell tests that meet both challenges by using mixtures of odorous molecules that subjects perceive as unfamiliar. We call the odor of these mixtures unfamiliar because subjects cannot readily describe what they are smelling. SMELL-S measures olfactory sensitivity, the ability to detect increasing dilutions of a mixture of odorants. SMELL-R is an olfactory resolution test that measures the ability of subjects to discriminate the smell of two mixtures that become progressively more similar as the test proceeds. Neither of the tests requires that subjects match words with a smell percept. We show that SMELL-S and SMELL-R are highly reliable olfactory tests that overcome problems with odor-specific insensitivity and that can be applied without adaptation to subjects in a different country. We expect that these tests, applied in combination, will provide the sensitivity and specificity required for early diagnosis of smell dysfunction in different populations.

Results

Designing SMELL-S and SMELL-R Smell Tests. To improve currently available diagnostic tools for testing olfactory function, we created two smell tests based on odorant mixtures. The Olfactory Sensitivity Test (SMELL-S) measures sensitivity to a mixture of 30 monomolecular odorants (Fig. 1*A*). The Olfactory Resolution Test (SMELL-R) measures the ability of subjects to discriminate the smell of pairs of such mixtures with an increase in overlapping components (24, 25) (Fig. 1*B*). Tests were presented in glass jars or vials as triangle tests, in which subjects were asked to pick out the stimulus with the strongest odor (SMELL-S) or the odd odor (SMELL-R). Both tests used adaptive staircase procedures that are standard in clinical olfactory testing (26) (Fig. 1).

Test–Retest Reliability. Effective diagnostic tests must be designed with high test–retest reliability. We therefore measured the reliability of SMELL-S and SMELL-R in a population of subjects with a self-reported normal sense of smell (experiment 1; Fig. 2*A*). We tested two versions of SMELL-S and SMELL-R (v1 and v2), which differed in the 30 components used for the mixtures. We also carried out conventional threshold tests with the monomolecular odorants phenylethyl alcohol and butanol. All tests were self-administered with stimuli presented in glass jars. We excluded data from the butanol threshold test because the stimulus was not stable throughout the testing period (Dataset S1).

To assess test–retest reliability for SMELL-S, we computed the absolute difference in test–retest scores for each subject (Fig. 2*B*).

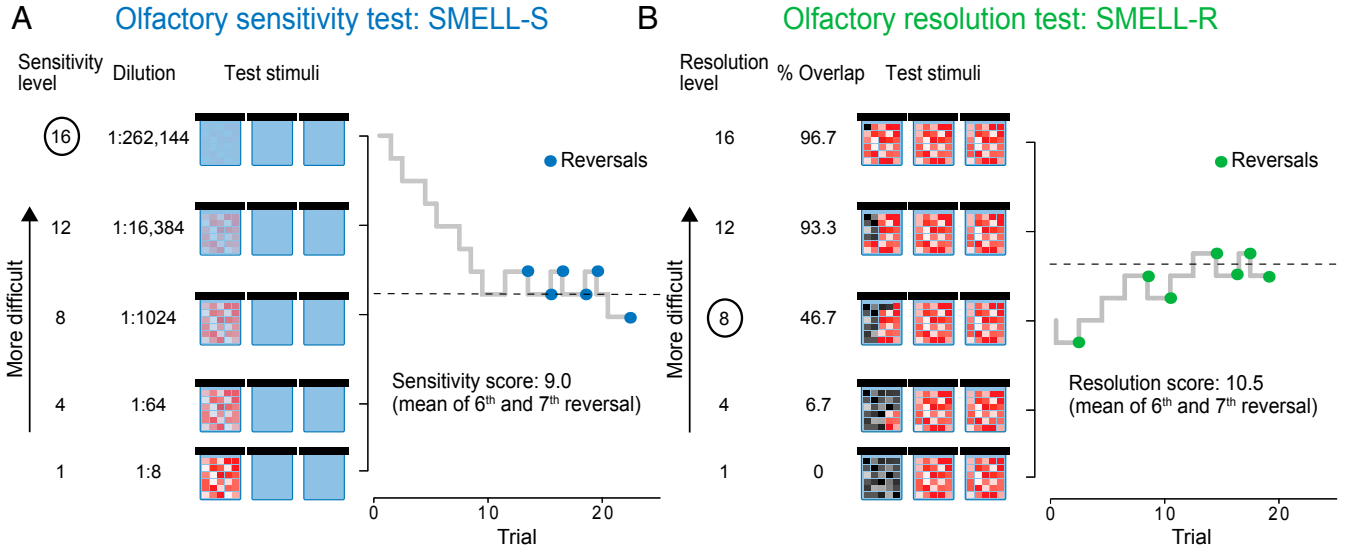

Fig. 1. SMELL-S olfactory sensitivity and SMELL-R olfactory resolution tests. (*A*) Schematic of triangle test stimuli for SMELL-S, comprising two glass vials containing solvent (blue) and one containing increasingly diluted mixtures of 30 molecules (red–white mosaic). Olfactory sensitivity of a subject measured with SMELL-S [Subject Expt 1-A023, SMELL-S (v2)]. (*B*) Schematic of triangle test stimuli for SMELL-R, comprising two jars containing the same mixture of 30 molecules (red–white mosaic) and one containing mixtures of 30 molecules (black–gray mosaic) with an increasing number of molecules shared with the other two. Olfactory resolution of a subject measured with SMELL-R [Subject Expt 1-A016, SMELL-R (v1)]. Circles in *A* and *B* indicate starting level for each test.

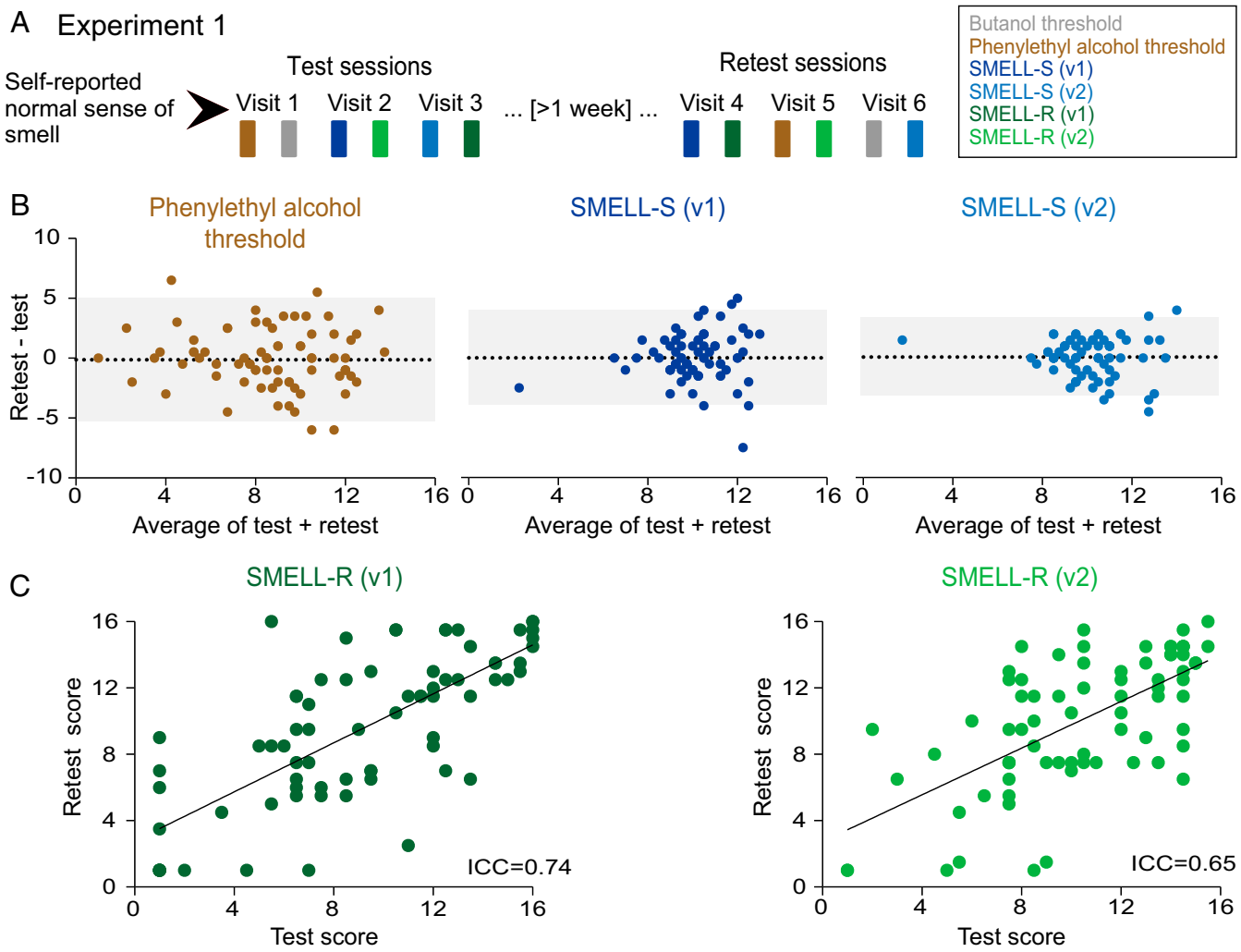

Fig. 2. Test–retest reliability and relationship between SMELL-S and SMELL-R tests. (*A*) Experiment 1 design, showing one example of the many different presentations of the six smell tests. SMELL-R tests were always administered after SMELL-S or threshold tests in a given visit. (*B*) Bland–Altman plots of the indicated tests where each dot represents data from one subject. Black dotted lines represent the average of differences between retest and test scores, and gray areas indicate 95% limits of agreement (average difference ±1.96 SD of the difference, $n = 74$–75). (*C*) Test and retest scores for SMELL-R (v1) and SMELL-R (v2) where each dot represents data from one subject (ICC, intraclass correlation coefficient; $n = 73$–75).

The bias, as defined by the difference between the average of the test and retest scores, was close to zero for all three tests. This indicates that subjects did not show systematically different performance between test and retest sessions. The 95% limits of agreement were smaller for the two SMELL-S tests than the phenylethyl alcohol threshold test (Fig. 2*B*). We did not calculate test–retest correlations (which can be calculated from Dataset S1) because of the lower interindividual variability of the two versions of SMELL-S compared with the phenylethyl alcohol threshold test. If the variability in a sample is low, the correlation coefficient tends to be low. Therefore, it is difficult to compare test–retest reliability when different tests have significantly different interindividual variability (27). The phenylethyl alcohol threshold test is commercially available as Sniffin' Sticks, a well-validated test administered by clinical staff that utilizes felt-tip pens for odorant delivery (26, 28). To confirm that our self-administered phenylethyl alcohol threshold test presented in glass vials produced results comparable to Sniffin' Sticks, we reinvited 23 subjects from experiment 1 and administered the Sniffin Sticks' version of the phenylethyl alcohol threshold test. There was a strong correlation between the phenylethyl alcohol threshold self-administered in glass vials and Sniffin' Sticks administered by a research assistant ($r = 0.87$; 95% confidence interval: 0.72–0.95, Pearson correlation). We conclude that SMELL-S is a reliable test of olfactory sensitivity.

We next examined the test–retest reliability of SMELL-R. Because the interindividual variability between SMELL-R (v1) (mean ± SD, 9.3 ± 4.3) and SMELL-R (v2) (mean ± SD, 10.2 ± 3.5) did not differ significantly ($P = 0.074$, F test) (Fig. 2*C*), we calculated the intraclass correlation coefficient (ICC) for the SMELL-R tests. By this metric, the two versions of SMELL-R are very reliable (Fig. 2*C*), and we selected SMELL-R (v2) for the remaining experiments in the study.

Addressing the Problem of Odor-Specific Insensitivity. Although the frequency of total loss of sensitivity to phenylethyl alcohol in healthy subjects is low, diminished sensitivity to this odorant is frequent, which may lead to misdiagnosis. This is illustrated by the large interindividual variability in sensitivity to this rose-like odor (20, 30). A test based on mixtures of components minimizes the impact of the sensitivity to any single component on the overall test score and thereby strengthens the diagnosis of general olfactory dysfunction. To explore how odor-specific sensitivity affects the accuracy of smell dysfunction diagnosis, we compared the performance of subjects in experiment 1 on smell

tests that used monomolecular stimuli or mixtures. The variability in test scores across all subjects in experiment 1 of the phenylethyl alcohol threshold was significantly higher than that of SMELL-S (v1) and SMELL-S (v2) (Fig. 3*A*). Of the seven subjects in the lowest 10th percentile in the phenylethyl alcohol threshold test, only one was in the lowest 10th percentile for both

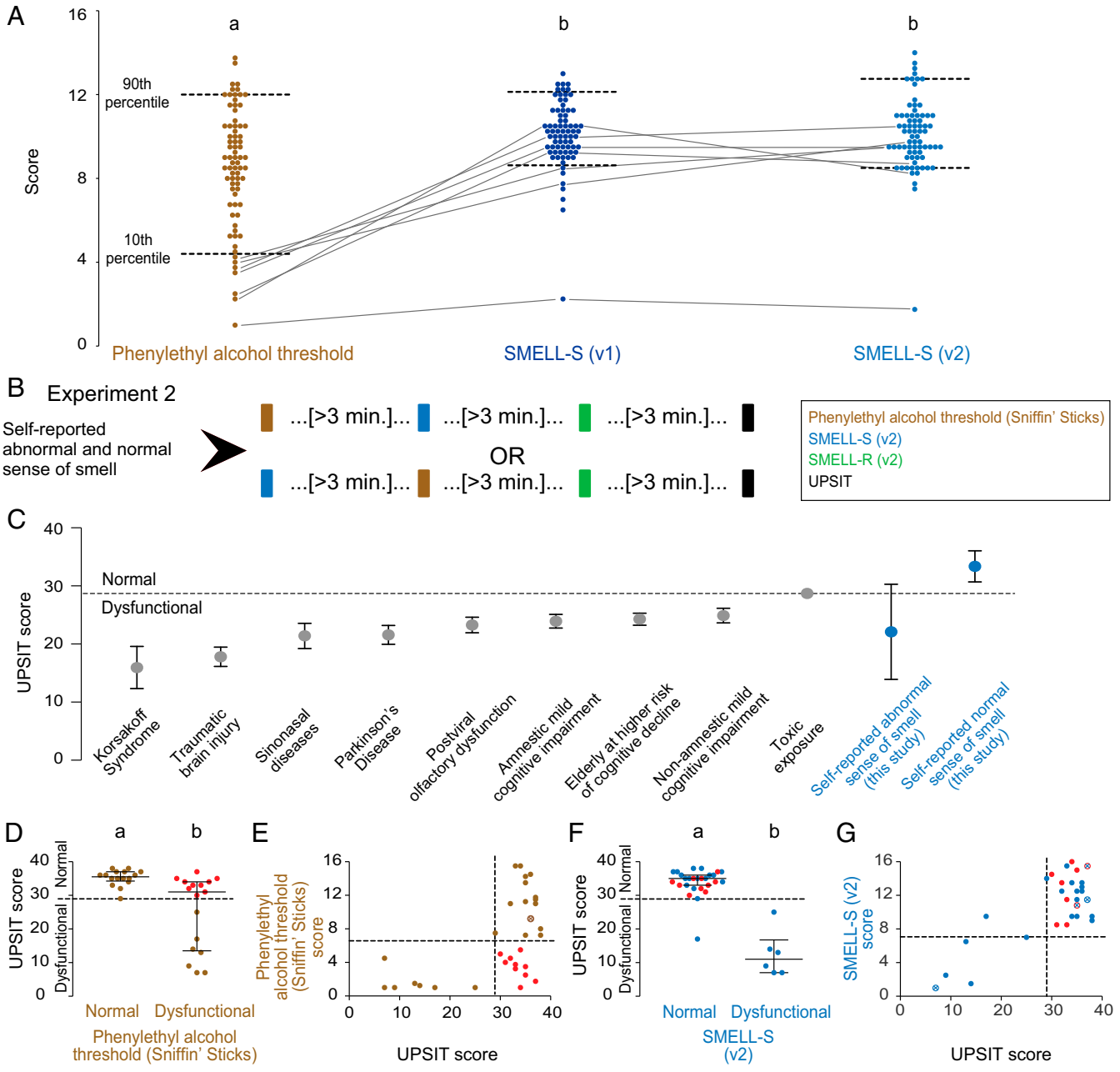

Fig. 3. Addressing the problem of odor-specific insensitivity. (*A*) Subject average scores of the indicated tests from experiment 1 between test and retest. Data marked with different letters (a or b) indicate significantly different interindividual variance between groups ($P < 0.0001$; Conover squared ranks test, $n = 74$–75). Data from the seven lowest scoring subjects in the phenylethyl alcohol test are connected by lines. (*B*) Experiment 2 design. (*C*) The relationship between different etiologies of olfactory dysfunction and UPSIT scores in published studies, as well as of experiment 2 subjects divided by self-reported smell abilities (mean ± 95% confidence interval). References: Korsakoff Syndrome (46), traumatic brain injury (48), sinonasal disease (48), Parkinson's disease (49), postviral olfactory dysfunction (48), amnestic mild cognitive impairment (9), elderly at higher risk of cognitive decline (9), nonamnestic mild cognitive impaired (9), and toxic exposure (50). (*D*) UPSIT scores for subjects divided into normal and dysfunctional groups according to phenylethyl alcohol threshold (Sniffin' Sticks) (cutoff score, 6.5) performance, where each dot represents data from one subject. Subjects scored as normal by the UPSIT but dysfunctional by Sniffin' Sticks phenylethyl alcohol test are colored in red in *D* and *E*. Data from the remaining subjects are colored brown here and in *E*. Data labeled with different letters (a or b) are significantly different ($P = 0.0003$, Mann–Whitney test). Medians and interquartile range are represented. (*E*) Comparison of UPSIT and phenylethyl alcohol (Sniffin' Sticks) scores. (*F*) UPSIT scores for subjects divided into normal and dysfunctional groups according SMELL-S (v2) (cutoff score, 7) performance, where each dot represents data from one subject. Subjects scored as normal by the UPSIT but dysfunctional by Sniffin' Sticks phenylethyl alcohol test in *D* are colored red in *F* and *G*, while the others are in blue. Data labeled with different letters (a or b) are significantly different ($P < 0.0001$; Mann–Whitney test) and represented as median and interquartile range. (*G*) Comparison between UPSIT and SMELL-S (v2) scores. Subjects with identical values in *E* and *G* are indicated by superimposed open circles and an X and retain the specified color coding.

versions of SMELL-S. Based on these results, we suspect that the six other subjects have specific insensitivity to phenylethyl alcohol rather than impaired olfactory function.

We next compared the SMELL-S test with the Sniffin' Sticks phenylethyl alcohol threshold test and the North American version of the UPSIT (experiment 2; Fig. 3*B*). Since SMELL-S (v2) had the narrowest 95% limits of agreement (Fig. 2*B*), we used this version of SMELL-S for the rest of this study. In experiment 2, we assessed the performance of subjects with a self-reported normal or abnormal sense of smell on the UPSIT, Sniffin' Sticks, and SMELL-S (v2). Based on results in Fig. 3*A*, we anticipated that SMELL-S (v2) would be more accurate than the Sniffin' Sticks phenylethyl alcohol threshold test in identifying subjects with smell dysfunction. We used the UPSIT to benchmark the performance of the Sniffin' Sticks phenylethyl alcohol threshold test compared with SMELL-S (v2). Because this smell test is composed of 40 different items, the final score is not strongly affected by odor-specific insensitivity to any given stimulus among the 40 items of the test. The UPSIT has many cutoffs for different degrees of olfactory dysfunction—for example, normosmia; mild, moderate, and severe microsmia; anosmia; malingering. These cutoff scores change according to gender and are influenced by age. For our study, we chose a single cutoff score based on disease etiology (Fig. 3*C*). Based on this analysis and consistent with an earlier study (3), we defined normal olfactory function as an UPSIT score of 29 and over and smell dysfunction as an UPSIT score of 28 and lower (Fig. 3*C*). In experiment 2, the mean score of the subjects with self-reported smell dysfunction was below this cutoff, whereas the mean score of those with self-reported normal sense of smell was above the cutoff (Fig. 3*C*). For the Sniffin' Sticks phenylethyl alcohol threshold test, we used the cutoff specified by the manufacturer, with normal defined as a score higher than 6.5 and dysfunctional a score of lower than 6.5.

Subjects in experiment 2 were divided into normal and dysfunctional according to their performance on the Sniffin' Sticks phenylethyl alcohol threshold test (Fig. 3*D*). If we use the UPSIT cutoff score of 29 as a metric of olfactory dysfunction, 10 subjects with a normal UPSIT score would have been diagnosed as having olfactory dysfunction by the Sniffin' Sticks phenylethyl alcohol threshold test (Fig. 3 *D*–*G*, red dots). When we divided subjects according to performance on the SMELL-S (v2) test using a cutoff score of 7 (Fig. 4), we found that only one subject was given a different diagnosis using the UPSIT than with SMELL-S (v2) (Fig. 3*F*).

Diagnostic Accuracy. To be useful in the clinic, diagnostic smell tests must correctly identify patients with smell dysfunction by

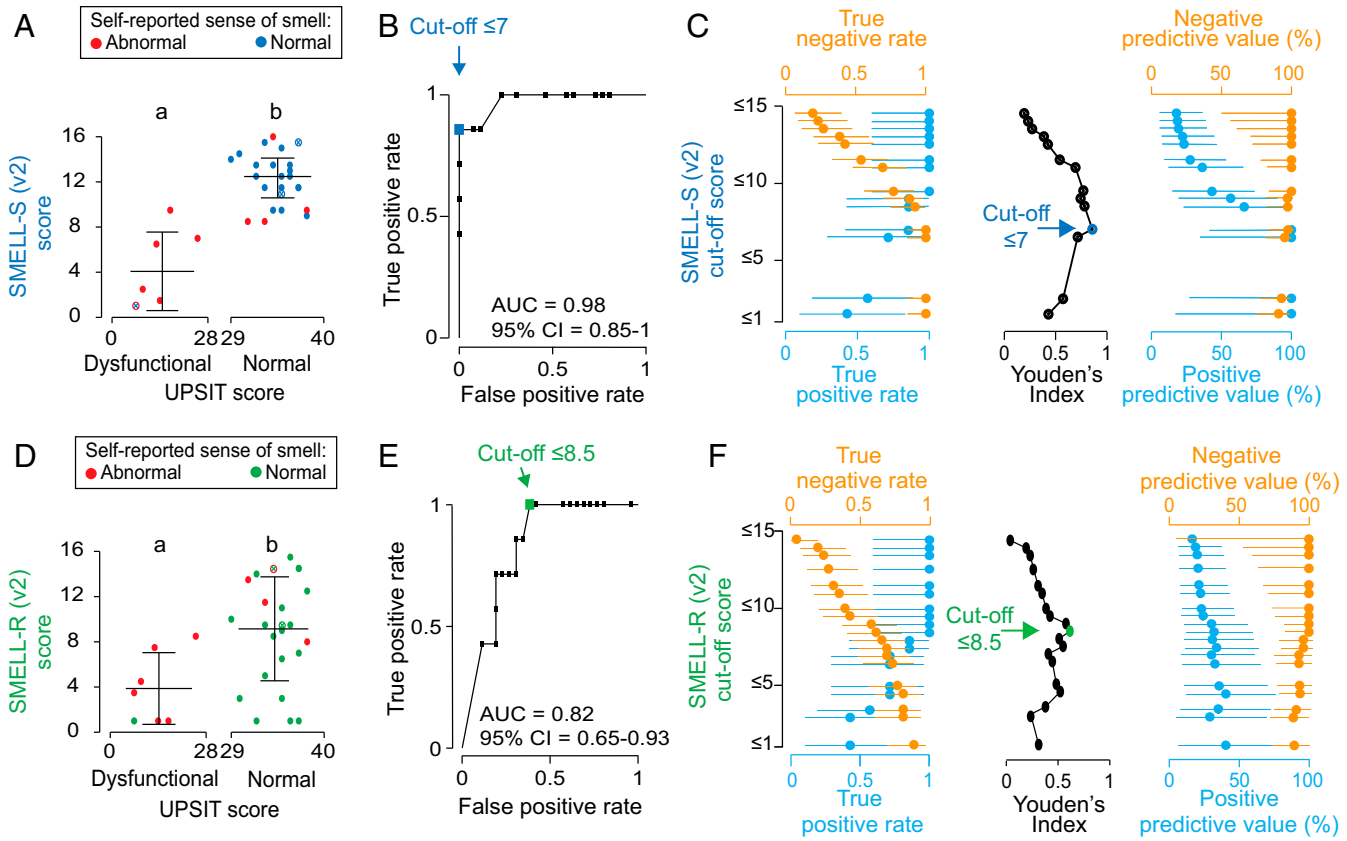

Fig. 4. SMELL-S and SMELL-R diagnostic accuracy. (*A*) Comparison of UPSIT and SMELL-S (v2) scores for experiment 2 subjects (mean ± SD). Subjects were divided into dysfunctional ($n = 7$) and normal ($n = 26$) using an UPSIT cutoff score of 29. Data labeled with different letters (a or b) are significantly different ($P = 0.0005$, two-sided unpaired t test with Welch's correction). (*B*) Area under the ROC curve (AUC) for SMELL-S (v2). The optimal cutoff is indicated by the blue square. (*C*) Plots of four measures of diagnostic accuracy resulting from different cutoff scores for SMELL-S (v2) (percentage ± 95% confidence interval). The optimal cutoff score for olfactory dysfunction defined by Youden's Index (*Center*) is indicated by the blue dot. (*D*) Comparison of UPSIT and SMELL-R (v2) scores for experiment 2 subjects (mean ± SD). Subjects were divided into dysfunctional ($n = 7$) and normal ($n = 26$) using an UPSIT cutoff score of 29. Data labeled with different letters (a or b) are significantly different ($P = 0.0035$, two-sided unpaired t test with Welch's correction). Subject Expt 2-A006 had a self-reported normal sense of smell but low scores on the UPSIT as well as SMELL-S (v2) in *A* and SMELL-R (v2) in *D*. (*E*) Area under the ROC curve (AUC) for SMELL-R (v2). The optimal cutoff is indicated by the green square. (*F*) Plots of four measures of diagnostic accuracy resulting from different cutoff scores for SMELL-R (v2) (rate or percentage ± 95% confidence interval). The optimal cutoff score for olfactory dysfunction defined by Youden's Index (*Center*) is indicated by the green dot. Subjects with identical values in *A* and *D* are indicated by superimposed open circles and an X and retain the specified color coding.

balancing false positive and false negative results. To establish a diagnostically optimal cutoff score for SMELL-S (v2) and SMELL-R (v2), we divided subjects into dysfunctional and normal using an UPSIT cutoff score of 29 and examined SMELL-S (v2) scores of self-reported normal and abnormal subjects in these two groups (Fig. 4*A*). Subjects with normal UPSIT scores had significantly

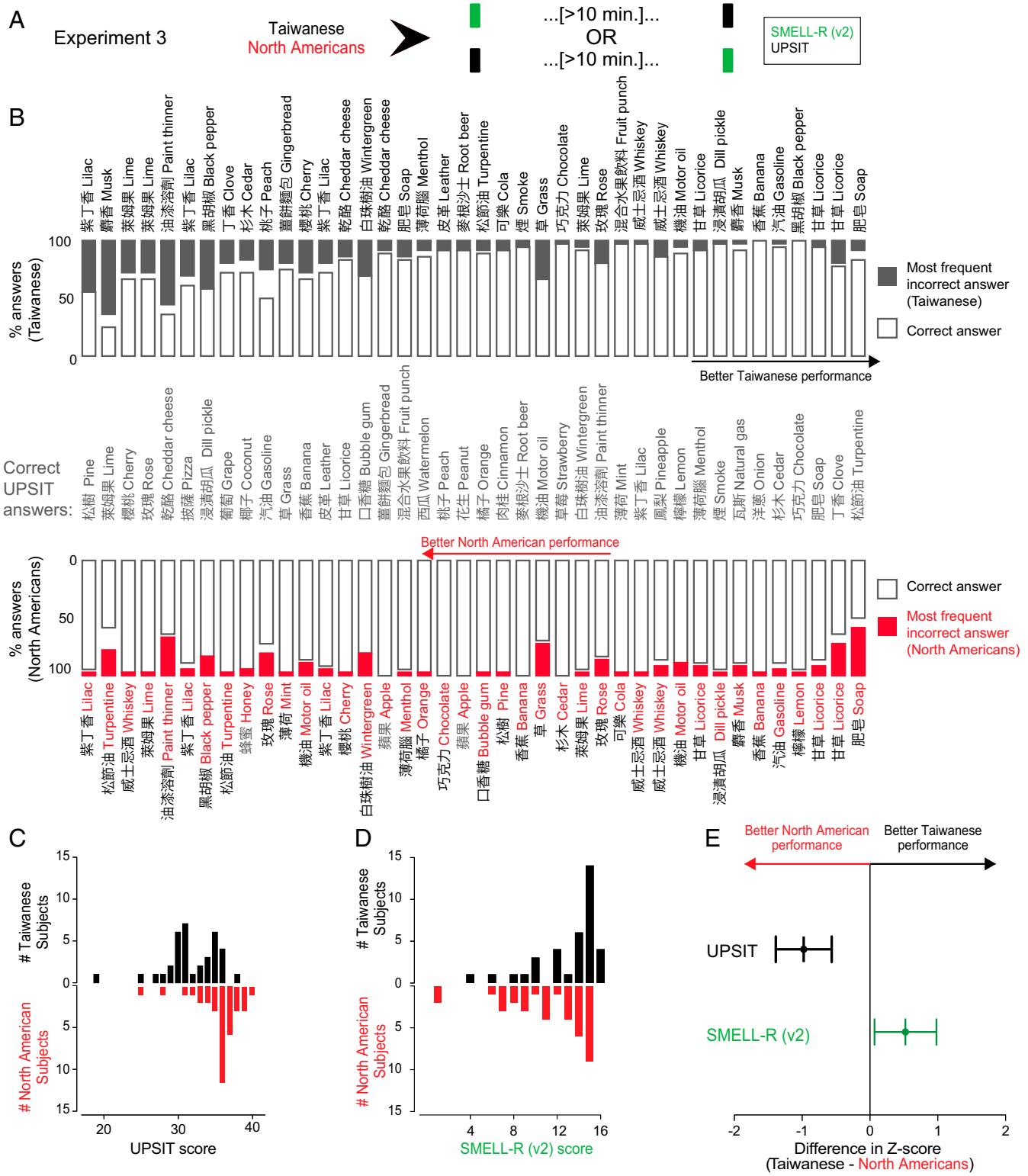

Fig. 5. Addressing the problem of smell testing in different populations. (*A*) Experiment 3 design. (*B*) Performance of Taiwanese ($n = 36$) (*Top*) and North American ($n = 36$) (*Bottom*) subjects for individual UPSIT items; open bar bars indicate correct answers, and solid bars indicate the most frequent incorrect answer. Data are sorted by the size of the difference between the two populations. (*C* and *D*) Histogram of North American and Taiwanese subject scores for the UPSIT (*C*) and SMELL-R (v2) (*D*). (*E*) Cross-population comparison of UPSIT and SMELL-R (v2) (mean ± 95% confidence interval) for subjects in *C* and *D*.

higher SMELL-S (v2) scores than those who were dysfunctional (Fig. 4*A*). Subjects with a self-reported normal sense of smell (blue dots in Fig. 4*A*) had significantly higher SMELL-S (v2) scores (median, 12.5; interquartile range, 11–14) than whose with a self-reported abnormal sense of smell (median, 7.75; interquartile range, 2.25–9.50) (red dots in Fig. 4*A*) ($P = 0.0011$, Mann–Whitney test).

We next determined the overall accuracy of SMELL-S (v2) and selected an optimal cutoff score to differentiate normal and dysfunctional subjects (Fig. 4 *B* and *C*). The standard measure of clinical test accuracy is the area under the receiver operating characteristic (ROC) curve, which plots the true and false positive rates at different cutoff scores. The area under the ROC curve of SMELL-S (v2) is 0.98 (95% confidence interval: 0.85–1.00) (Fig. 4*B*), which is close to the perfect accuracy of 1.

To select the cutoff value for SMELL-S (v2) that optimally distinguishes normal and dysfunctional subjects, we calculated Youden's Index (31) at each of 14 SMELL-S (v2) cutoff scores. A Youden's Index value of 1 indicates no false positives and no false negatives (Fig. 4*C*). Based on this analysis, we suggest that the administration of SMELL-S (v2) with a cutoff value of 7 will be optimal to diagnose patients with olfactory dysfunction.

We carried out the same procedure to determine the accuracy of the SMELL-R olfactory resolution test. Subjects classified as dysfunctional by their UPSIT score had lower SMELL-R (v2) scores, and subjects classified as normal by UPSIT performance had higher SMELL-R (v2) scores (Fig. 4*D*). The area under the ROC curve for SMELL-R (v2) was 0.82 (95% confidence interval: 0.65–0.93) (Fig. 4*E*). The optimal cutoff assessed by Youden's Index was 8.5 (Fig. 4*F*). These proof-of-principle data show that SMELL-R at a cutoff value of 8.5 may be clinically useful for diagnosing smell dysfunction.

Addressing the Problem of Different Prior Olfactory Experiences. A major goal of this study was to develop a test that does not have to be adapted to different populations. To ask if SMELL-R (v2) performs well in different countries, we compared SMELL-R (v2) performance between Taiwanese and North American subjects (experiment 3; Fig. 5*A*). As a positive control, we used the North American version of the UPSIT for both populations, because previous work has shown that Taiwanese subjects have systematically lower scores on this test due to unfamiliarity with several of the test items (21). To enable self-administration of the UPSIT, we supplied Taiwanese subjects with a Chinese translation of the English multiple-choice questions in the test booklet. SMELL-R (v2) did not require any language translation because it is nonsemantic.

As expected, North Americans performed better on most of the items in the UPSIT, with the biggest differences found for "pine," "lime," "cherry," and "rose" (Fig. 5*B*). Even so, several items were frequently mistaken by Taiwanese subjects, including "paint thinner" when the correct answer was "cheddar cheese," "musk" instead of "lime," and "wintergreen" instead of "bubble gum." The North American subjects also struggled with the "cheddar cheese" item, also frequently mistaking it for "paint thinner", but in addition mistook "turpentine" for "soap," "motor oil" for "grass," and "clove" for "licorice." The overall UPSIT scores for Taiwanese subjects were significantly lower than those of the North American subjects (Fig. 5*C*) ($P < 0.0001$, Mann–Whitney test). In contrast, Taiwanese subjects scored higher on SMELL-R (v2) than the North American subjects (Fig. 5*D*) ($P = 0.0157$, Mann–Whitney test). The difference between the two populations was much smaller for SMELL-R (v2) than the UPSIT, as determined by calculating the difference in z scores (Fig. 5*E*). While we do not know the underlying cause for the superior performance of Taiwanese subjects on SMELL-R (v2), the results show that our test avoids the bias seen for the UPSIT, in which test performance was systematically higher in the population for which the test was developed. We conclude that SMELL-R (v2) can be applied across different populations without the need to adapt it to the local culture and language.

Discussion

In this study, we addressed current limitations in clinical testing for olfactory dysfunction by developing effective smell tests that overcome issues with odor-selective insensitivity and that can be utilized with different populations across the world.

The first objective of this work was to eliminate the problems inherent in olfactory sensitivity tests that rely on a single molecule. Although it is well known that specific insensitivity to individual odorant molecules is common in normal human subjects (17–19), commercial threshold tests use monomolecular stimuli such as butanol or phenylethyl alcohol to test olfactory sensitivity (26, 32). Our data suggest that this approach confounds specific and general olfactory sensitivity. We show here that the solution to this problem is to use mixtures of molecules instead of single molecules. We and other authors have shown that the inter- and intraindividual variability in threshold scores was reduced, and test–retest reliability was increased by testing olfactory sensitivity with odor mixtures rather than single molecules (33, 34). One previous study compared thresholds for single molecules to those for mixtures of 3, 6, or 12 components and concluded that the intra- and interindividual variability of the threshold decreases with increasing number of molecules in the mixture (33). A recent study came to a similar conclusion, comparing the threshold for phenylethyl alcohol to the threshold for a mixture of three molecules (34).

Although the rate of specific anosmia to phenylethyl alcohol is low, interindividual variability in sensitivity to this molecule is large (20, 30). It follows that diminished sensitivity could lead to false positive results, and therefore misdiagnosis. The SD we found for phenylethyl alcohol in 75 healthy subjects was 2.75, which is consistent with previous studies that reported SDs of 2.88 (30) and 2.78 (20). SMELL-S had much lower variability, with an SD of 1.6 for SMELL-S v1 and 1.7 for SMELL-S v2. We conclude that SMELL-S is a reliable, accurate, and effective method for measuring olfactory function without conflating general loss of smell sensitivity and specific insensitivity to an odorant.

A second objective of this project was to introduce a test of olfactory resolution (SMELL-R) that quantifies olfactory discrimination ability. Auditory and visual stimuli used in the clinic differ by tone frequency or letter size, leading to quantitative and standardized diagnostic tests such as the audiogram and the eye chart. In olfaction, it is more complicated to quantify similarity between olfactory stimuli. Currently available discrimination tests consist of several pairs of odorants that must be discriminated by the patient. There currently is no method to quantify how difficult the individual discrimination tests are. Is distinguishing "rose" from "leather" more or less difficult than discriminating "pineapple" and "licorice"? To overcome this problem, we used a physical scale based on the number of shared components between two mixtures. The more components two mixtures share, the more difficult it is to discriminate them (25). By using this physical scale, a patient's olfactory resolution can be reliably determined.

A third objective was to develop smell tests that utilize stimuli that have not been previously encountered by patients to minimize the influence of cultural and personal differences in prior olfactory experiences on the test results (22, 23). We accomplished this by using mixtures of 30 different molecules. These exact mixtures are very unlikely to be encountered outside the laboratory and are perceived as unfamiliar smells. Furthermore, before mixing them, the chemicals were diluted so that they had approximately equal intensity to ensure that the percept of the mixture is not dominated by a single odorant. The resulting smells of such mixtures have been described as "olfactory whites" (24). Using these stimuli is an

improvement over the use of odorants that can be readily linked to their usual source but only by those who have prior experience with it (21, 35, 36).

The proof-of-principle results with SMELL-S and SMELL-R presented here suggest that these tests will be useful in diagnosing smell dysfunction, but it is important to note that our sample sizes were comparatively small. Future studies with larger groups of patients with known olfactory dysfunction will be necessary to fully validate the tests. It will also be necessary to formulate the tests in a compact delivery system that automatically delivers stimuli and records subject responses. Modern advances in digital technology for odor delivery and data capture will enable this goal. Moving from these initial studies to a standardized clinical test will need to take into account the optimal solvents to assure odor stability (37) and the effect that the delivery system has on test performance (38). Finally, although the prototype test discussed here was self-administered by healthy volunteers with minimal training in about 30 min, we recognize that the use of SMELL-S and SMELL-R in geriatric patients, especially those suffering from neurodegenerative disease, will require further adaptation. Developing a universal olfactory test to reliably diagnose smell dysfunction is of great clinical importance not only because of the negative effects of smell dysfunction on quality of life but because olfactory dysfunction is frequent, can be clinically managed, and may be an effective biomarker for predicting Alzheimer's disease, Parkinson's disease, and other neurodegenerative diseases (9, 39).

Materials and Methods

General, Subjects. All behavioral testing with human subjects took place between March 2015 and December 2016 and was approved and monitored by the Institutional Review Board of The Rockefeller University in New York, except the Taiwanese arm of experiment 3, which was approved by the Institutional Review Board of Taichung Veterans General Hospital in Taichung, Taiwan. North American subjects were recruited by The Rockefeller University Clinical Research Recruitment and Outreach Support Service (40). Taiwanese subjects were recruited by the nursing staff of the Department of Otorhinolaryngology at the Taichung Veterans General Hospital (Taiwan). All subjects gave their written informed consent to participate in these experiments and were compensated for their time. All North American and Taiwanese subjects were able to understand and follow instructions in English or Mandarin, respectively. Subjects were aged 18 or over and agreed to refrain from using perfume or cologne and ingesting anything except water 1 h before the study visit. At the beginning of each visit, subjects washed their hands with fragrance-free soap. For subjects reporting a normal sense of smell and taste, we excluded subjects who presented with current or past history of conditions that might be related to smell loss (acute or chronic rhinosinusitis, nasal tumor, upper respiratory tract infection or head trauma that altered the sense of smell for more than 1 mo, history of brain or sinonasal surgery, asthma, stroke, neurodegenerative disease, radiation therapy or chemotherapy, active smoking, or consumption of medication affecting the sense of smell during the study). Participants with self-reported smell dysfunction were not subject to these exclusion criteria. All raw data in the paper, including details about the demographics of the subjects, odorants, and composition of the test stimuli are in Dataset S1.

General, Tests. To allow for self-administration and automatic data collection, we designed a custom computer application that was used for the phenylethyl alcohol and butanol threshold tests and also the SMELL-S and SMELL-R tests. The testing station comprised a computer, wireless mouse, barcode scanner, and trays with numbered stimulus containers labeled with bar codes. Triangle tests were set up so that subjects were never tested with the same set of stimuli twice in a row, to avoid the situation where subjects remembered their answers from the previous trial. Subjects used a barcode scanner to register test data automatically. Subjects took between 20 and 35 min to complete each smell test, with the exception of the UPSIT, which took 10–15 min. A standard intertrial interval was imposed to avoid odor adaptation by requiring subjects to play a computer game for 20 s.

SMELL-S and SMELL-R were created with four different mixtures of 30 molecules drawn from a panel of 109 monomolecular, intensity-matched chemicals. These odorants were selected from stimuli utilized in previous psychophysical studies (24, 41). We used only molecules that minimally activated the trigeminal system, because such stimuli can be detected by anosmic subjects (42, 43). A characteristic of trigeminal activation by a molecule is a fresh, cold, burning, eucalyptus, pungent, or tickling sensation. We used a lateralization task in which an odorant is applied into only one nostril to assign a lateralization score to each molecule. It is possible to localize the stimulated nostril if it activates the trigeminal system. In contrast, it is much harder to localize an olfactory stimulus (44). Lateralization tasks were self-administered by one investigator. Two disposable squeeze bottles were placed in a device facilitating simultaneous squeezing and stimulus delivery in each nostril. Only one bottle was filled with an odor stimulus. The tip of each bottle was fitted with a foam piece that conformed to the investigator's nostril and was placed at the entrance of each nostril. The investigator squeezed both bottles simultaneously and attempted to localize which nostril had received the stimulus. After each task, the device was spun on a rotating platform to randomize the odor-stimulus side. The final score corresponded to the number of correct tasks. There were a total of 20 tasks (45). As a control experiment, we found that the lateralization score of the trigeminal stimulus eucalyptol [PubChem compound identification (CID): 2758] at pure concentration was high (median, 20; interquartile range, 19.25–20; four trials). The lateralization score of the olfactory stimulus vanillin (CID: 1183) at pure concentration was low (median, 6.5; interquartile range, 5–12.5; six trials). The difference between the lateralization scores of eucalyptol and vanillin was statistically significant ($P = 0.0009$, Mann–Whitney test). Each candidate for the mixtures was tested once. We included candidates with a score of 11 and below in the design of the mixtures (Dataset S1).

To intensity-match molecules to be used in mixtures, odorants were diluted and three investigators individually classified them as "too weak," "well matched," or "too strong." The concentration of too weak stimuli was increased and that of too strong stimuli decreased by a factor of 10. Weak components that could not be intensity-matched even at pure concentrations were excluded from the pool of odorants. We repeated this process until most of the components fell into the optimal intensity range. For 18 components investigators could not reach a consensus about intensity, but these were nevertheless used in the mixtures (CID: 1068, 7969, 31244, 9589, 17898, 104721, 3314, 14491, 62144, 7583, 7983, 60999, 251531, 7799, 61151, 9609, 8118, and 89440). With these components, we created four mixtures of 30 components. The SMELL-S (v1) mixture was used as the ODD odor in SMELL-R (v1), and the SMELL-S (v2) mixture was used as the CONTROL odor in SMELL R (v2). The mixtures for SMELL-R (v1) CONTROL odor and SMELL-R (v2) ODD odor were unique to these tests. Details of all mixtures are in Dataset S1.

Stimuli for the threshold tests and SMELL-S were presented to subjects with amber glass vials (height, 95 mm; diameter, 28 mm). Stimuli for SMELL-R were presented to subjects with amber glass jars (height, 51 mm; diameter, 55 mm). The complete list of stimuli used in this study is in Dataset S1.

Threshold Tests: Phenylethyl Alcohol and Butanol. Threshold tests were administered as a series of triangle tests. Subjects were presented with three vials: two contained 1 mL solvent (paraffin oil) and one contained either phenylethyl alcohol or butanol diluted in solvent in a total volume of 1 mL. Tests comprised 16 different concentrations generated by serial dilutions (1:2) of either odorant in paraffin oil, with the starting concentrations at 0.0313% for phenylethyl alcohol and 0.25% for butanol. The subject was prompted to sniff each vial and select the one with the strongest perceived odor using an adaptive staircase procedure commonly used in smell testing (26). If they were unable to detect any difference among the three vials, they were prompted to choose one at random. The procedure started at the lowest concentration. If they identified an incorrect vial, the second next higher concentration was presented and so on, until they identified the correct vial. If the subjects identified the correct vial, they were retested at the same concentration. If they identified the correct vial in this retest, they were tested at the next lower concentration. If they identified an incorrect vial, they were tested at the next higher concentration. A reversal is when the direction in which the concentration is changed reverses. The procedure ended after the seventh reversal, or after the subject failed the level with the highest concentration twice in row, or succeeded with the lowest concentration level five times in row. The threshold was defined as the average of the concentrations at which the last two reversals occurred. If the highest concentrations were not correctly identified twice, the score was 1. If the lowest was identified five times in a row, the score was 16.

SMELL-S Olfactory Sensitivity Test (v1 and v2). For SMELL-S (v1) and SMELL-S (v2), we prepared 19 serial dilutions in paraffin oil (1:2) of two different mixtures of 30 monomolecular odorants and used the last 16 dilutions, such that the tests ranged from easiest (level 1, 1:8 dilution) to most difficult (level 16, 1:262,144 dilution). Subjects were asked to sniff three vials, one of which

was filled with 1 mL of a mixture of 30 components and the other two were filled with 1 mL of solvent (paraffin oil). Subjects were instructed to pick out the one vial with the strongest perceived odor. If they were unable to detect any difference among the three vials, they were prompted to choose one at random. The procedure started at the lowest concentration (level 16). We calculated the SMELL-S sensitivity score following the same adaptive staircase procedure described above. For each subject, we measured the olfactory sensitivity with two versions of the test, SMELL-S (v1) and SMELL-S (v2), which differed only by the chemical composition of the mixtures.

SMELL-R Olfactory Resolution Test (v1 and v2). For SMELL-R (v1) and SMELL-R (v2), we prepared 16 pairs of mixtures of 30 monomolecular odorants that differed in how many components the two mixtures in the pair share from 0% (easiest; level 1) to 96.7% (most difficult; level 16). To create 16 levels of increasing overlapping components, we progressively replaced components of a mixture of 30 molecules (we termed this the ODD odor) with components from another mixture of 30 components that did not change in composition across the levels (we termed this the CONTROL odor). Increasing the level of difficulty by one point corresponds to an addition of two overlapping molecules between both mixtures, except from level 15–16, where we added only one shared molecule. Stimuli (8 mL) were introduced into jars containing absorbent cotton pads. Subjects were asked to sniff the contents of three jars, one of which was filled with 8 mL of a mixture of 30 components and the other two were filled with 8 mL of a mixture of 30 components with different degrees of overlap with the first jar. Subjects were instructed to pick out the odd jar. If they were unable to detect any difference among the three jars, they were prompted to choose one at random. Triangle tests started at a medium difficulty (level 8). If they identified the incorrect jar, the next easier level was presented. We calculated the SMELL-R resolution score following the same adaptive staircase procedure described above. For each subject, we measured the olfactory resolution with two versions of the test, SMELL-R (v1) and SMELL-R (v2), which differed only in the chemical constituents of the two sets of mixtures.

Sniffin' Sticks Phenylethyl Alcohol Threshold Test. The Sniffin' Sticks (26) threshold phenylethyl alcohol threshold test is a commercial product that uses felt-tip pens filled with odorant instead of ink for odor presentation. In this study, we used threshold module (2-phenyl ethanol) of the extended Burghart Sniffin' Sticks test (item LA-13-00015; Burghart Messtechnik). The test comprises pens containing 16 serial dilutions of phenylethyl alcohol (1:2) in solvent (propylene glycol) with a starting concentration of 4%. The test was administered as a triangle test. Three pens were presented to the subjects by the investigator in a randomized order. Two pens contained the solvent only, and the third pen contained the diluted odorant. Subjects were blindfolded with a disposable mask because the color code of the Sniffin' Sticks reveals which pen contains the odor, and subjects were asked to identify the pen with the strongest perceived odor. The procedure started at the lowest or second lowest concentration of odorant (level 16 or 15, respectively). We calculated the threshold score following the same adaptive staircase procedure described above except that the threshold was defined as the average of the last four reversals.

UPSIT. The UPSIT (marketed as the Smell Identification Test by Sensonics International) is a well-validated and self-administered smell identification test widely used in the United States (46). The test consists of four different 10-page booklets, with a total of 40 stimuli. On each page, there is a different "scratch and sniff" strip that is coated with a microencapsulated odorant and four words to choose from to describe the smell. Subjects used the tip of a pencil to release the smell of the stimuli. Subjects sniffed the odorant and selected the one word among the four options (for example, "paint thinner," "cherry," "coconut," or "cheddar cheese") that most closely matched their perception of the smell. Subjects entered their answers to the 40 multiple-choice questions manually into a booklet, and investigators transferred the data manually into a spreadsheet. UPSIT performance was scored as the number of correct answers out of 40. We used the same North American UPSIT (46) on subjects at Rockefeller University and Taichung Veterans General Hospital. The Taiwanese subjects were given a reference sheet on which the English multiple-choice questions in the UPSIT booklets were translated into Chinese by R.-S.J. (21) (Fig. 5*B*).

Experiment 1, Design. In this protocol (Rockefeller University IRB Protocol JHS-0862), we studied the test–retest reliability of SMELL-S and SMELL-R. We invited volunteers with self-reported normal sense of smell and taste to the Rockefeller University Hospital for six visits (Fig. 2*A*). During these six visits, six olfactory tests were performed, each of them once during a test session (visit 1–3) and then again during a retest session (visit 4–6). There was a gap of at least 1 wk between the last test visit (visit 3) and the first retest visit (visit 4) and a gap of at least 24 h between each of the other visits. At each visit, two of the six tests were performed. Although the order of the tests was randomized, in any visit where SMELL-R tests were administered, they were always administered after the SMELL-S or the threshold tests. This experiment was done between March and June 2015.

Experiment 1, Subjects. Seventy-five subjects (43 female) participated in this experiment, with a mean age of 44 (range, 21–74). Thirty-four subjects self-identified as White, 26 as Black, 6 as Asian, 2 as mixed race, and 7 as Other. Eleven subjects self-identified as Hispanic. It took an average of 21 d (range, 14–38 d) for subjects to complete all six visits in this experiment.

Experiment 1, Statistical Analysis. The ICC was used to measure absolute agreement between test and retest measures for the whole cohort. A sample of $n = 75$ subjects provided 95% confidence that the ICC in the population was larger than 0.67 based on a sample distribution that is centered on 0.8 (47). Bland–Altman plots were used as an auxiliary tool if significant differences in interindividual variability were found between compared tests (27) (Fig. 2*B*). We used the nonparametric Conover squared ranks test to assess equality of variance across threshold tests. Statistical significance was reached when $P < 0.05$ (Fig. 3*A*).

Experiment 2, Design. This experiment was carried out under Rockefeller University IRB Protocol JHS-0922 and was designed to evaluate the accuracy of our tests and whether SMELL-S can distinguish between subjects with specific anosmia to phenylethyl alcohol but an otherwise normal sense of smell and subjects with smell dysfunction. During a single visit in December 2016, subjects performed four smell tests. The first two tests were either SMELL-S (v2) or the Sniffin' Sticks phenylethyl alcohol threshold test. The order of these first two tests was randomized. It was followed by SMELL-R (v2) and finally the UPSIT, as a validated commercial reference test. The investigators enforced a break of at least 3 min between tests. During some of the breaks, participants filled out a questionnaire to provide demographic information and answer questions about their sense of taste and smell (Dataset S1). In seven cases in the UPSIT tests in experiment 2, subjects did not provide an answer to a given item, and this was scored as an incorrect answer. The missing data correspond to three subjects who missed one item each and two subjects who missed two items each.

Experiment 2, Subjects. This experiment included 33 subjects (22 female), with a mean age of 48 (range, 21–76). Seventeen subjects self-identified as White, eight as Black, three as Asian, two as mixed race, one as other. Two subjects opted out of self-reporting race. Four subjects self-identified as Hispanic. We re-enrolled 23 subjects from experiment 1 who self-reported a normal sense of smell and taste. These 23 were selected based on their threshold test results to have approximately even representation of subjects with low, medium, and high sensitivity to phenylethyl alcohol. In addition, we recruited 10 subjects with self-reported smell dysfunction. The self-reported etiologies are reported in Dataset S1.

Experiment 2, Statistical Analysis. We performed a power analysis and determined that a study with 32 subjects (8 with smell loss and 24 with a normal sense of smell) guarantees 80% power at 5% significance to detect an area under the ROC curve greater than 0.78. Since our actual study included 33 subjects, we carried out a post hoc power analysis using the parameters above to show that we can detect an area under the ROC curve greater than 0.79. We employed Youden's Index (31) to find the best cutoff score for SMELL-S and SMELL-R to maximize correct classification of the olfactory sensitivity and resolution of a subject, respectively (Fig. 4 *C* and *F*). We used two-sided unpaired *t* test with Welch's correction to test for differences between SMELL-S and SMELL-R score in normal and dysfunctional groups (Fig. 4 *A* and *D*).

Experiment 3, Design. In this experiment, we investigated how SMELL-R performs on different populations by comparing Taiwanese Taichung Veterans General Hospital (IRB Protocol TCVGH CE16119B) and North American (Rockefeller University IRB Protocol JHS-0901) subjects. The North American subjects were tested at The Rockefeller University Hospital, and the Taiwanese subjects were tested in the Department of Otolaryngology at Taichung Veterans General Hospital. The experimental design was the same in both institutions. Each subject came to the test site for a single visit, during which

subjects performed the SMELL-R (v2) and UPSIT, separated by a 10-min break, in randomized order (Fig. 5*A*).

Experiment 3, Subjects. Thirty-six subjects were recruited at both sites. All subjects were born and raised in their respective country, had never traveled to the opposite country, and had a self-reported normal sense of smell and taste. In the North American group, the mean age was 25 (range, 19–30), 23 of 36 subjects were female, and 8 self-identified as White, 14 as Black, 4 as Asian, 9 as mixed race, and 1 as American Indian or Alaska native. Six self-identified as Hispanic. In the Taiwanese group, the mean age was 26 (range, 19–30), and 26 of 36 subjects were female. Although we recruited subjects with a self-reported normal sense of smell, two of the North American subjects had UPSIT and SMELL-R (v2) scores below the cutoff for olfactory dysfunction (Fig. 5 *C* and *D*).

Experiment 3, Statistical Analysis. We used the unpaired *t* test with Welch's correction to test for differences in smell test performance between North American and Taiwanese subjects (Fig. 5 *C* and *D*).

Statistical Analysis. Normality of data were tested throughout using the Kolmogorov–Smirnov test, and the appropriate statistics were used according to the distribution of the data. SPSS (IBM) and Prism (Graphpad) were used for all statistical analyses.

ACKNOWLEDGMENTS. We thank our research volunteers for their time and interest in the study and the staff of The Rockefeller University Hospital Outpatient Clinic for invaluable support. Chris Vancil provided custom programming for the Rockefeller University Smell Study smell test computer interface, and Joel M. Correa da Rosa and Caroline Jiang provided expert biostatistical guidance. Yuanbo Wang provided a script to compute test scores in experiment 1. We thank Barry Coller, Ashutosh Kacker, Kevin Lee, and members of the L.B.V. laboratory for discussion and comments on the manuscript. This work was funded by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH), and Clinical and Translational Science Award (CTSA) Program UL1 TR000043. L.B.V. is an investigator of the Howard Hughes Medical Institute.

1. Brämerson A, Johansson L, Ek L, Nordin S, Bende M (2004) Prevalence of olfactory dysfunction: The Skövde population-based study. *Laryngoscope* 114:733–737.
2. Vennemann MM, Hummel T, Berger K (2008) The association between smoking and smell and taste impairment in the general population. *J Neurol* 255:1121–1126.
3. Liu G, Zong G, Doty RL, Sun Q (2016) Prevalence and risk factors of taste and smell impairment in a nationwide representative sample of the US population: A cross-sectional study. *BMJ Open* 6:e013246.
4. Hummel T, Nordin S (2005) Olfactory disorders and their consequences for quality of life. *Acta Otolaryngol* 125:116–121.
5. Keller A, Malaspina D (2013) Hidden consequences of olfactory dysfunction: A patient report series. *BMC Ear Nose Throat Disord* 13:8.
6. Leopold DA, Hornung DE, Schwob JE (1992) Congenital lack of olfactory ability. *Ann Otol Rhinol Laryngol* 101:229–236.
7. Karstensen HG, Tommerup N (2012) Isolated and syndromic forms of congenital anosmia. *Clin Genet* 81:210–215.
8. Temmel AF, et al. (2002) Characteristics of olfactory disorders in relation to major causes of olfactory loss. *Arch Otolaryngol Head Neck Surg* 128:635–641.
9. Devanand DP, et al. (2015) Olfactory deficits predict cognitive decline and Alzheimer dementia in an urban community. *Neurology* 84:182–189.
10. Brookmeyer R, Gray S, Kawas C (1998) Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset. *Am J Public Health* 88: 1337–1342.
11. Wilson RS, et al. (2009) Olfactory impairment in presymptomatic Alzheimer's disease. *Ann N Y Acad Sci* 1170:730–735.
12. Sindhusake D, et al. (2001) Validation of self-reported hearing loss. The Blue Mountains Hearing Study. *Int J Epidemiol* 30:1371–1378.
13. Landis BN, Hummel T, Hugentobler M, Giger R, Lacroix JS (2003) Ratings of overall olfactory function. *Chem Senses* 28:691–694.
14. Knaapila A, et al. (2008) Self-ratings of olfactory function reflect odor annoyance rather than olfactory acuity. *Laryngoscope* 118:2212–2217.
15. Sorokowska A, Drechsler E, Karwowski M, Hummel T (2017) Effects of olfactory training: A meta-analysis. *Rhinology* 55:17–26.
16. Nguyen DT, Rumeau C, Gallet P, Jankowski R (2016) Olfactory exploration: State of the art. *Eur Ann Otorhinolaryngol Head Neck Dis* 133:113–118.
17. Amoore JE (1967) Specific anosmia: A clue to the olfactory code. *Nature* 214:1095–1098.
18. Bremner EA, Mainland JD, Khan RM, Sobel N (2003) The prevalence of androstenone anosmia. *Chem Senses* 28:423–432.
19. Keller A, Hempstead M, Gomez IA, Gilbert AN, Vosshall LB (2012) An olfactory demography of a diverse metropolitan population. *BMC Neurosci* 13:122.
20. Zernecke R, et al. (2011) Correlation analyses of detection thresholds of four different odorants. *Rhinology* 49:331–336.
21. Jiang RS, et al. (2010) A pilot study of a traditional Chinese version of the University of Pennsylvania Smell Identification Test for application in Taiwan. *Am J Rhinol Allergy* 24:45–50.
22. Rabin MD (1988) Experience facilitates olfactory quality discrimination. *Percept Psychophys* 44:532–540.
23. Jehl C, Royet JP, Holley A (1995) Odor discrimination and recognition memory as a function of familiarization. *Percept Psychophys* 57:1002–1011.
24. Weiss T, et al. (2012) Perceptual convergence of multi-component mixtures in olfaction implies an olfactory white. *Proc Natl Acad Sci USA* 109:19959–19964.
25. Bushdid C, Magnasco MO, Vosshall LB, Keller A (2014) Humans can discriminate more than 1 trillion olfactory stimuli. *Science* 343:1370–1372.
26. Hummel T, Sekinger B, Wolf SR, Pauli E, Kobal G (1997) 'Sniffin' sticks': Olfactory performance assessed by the combined testing of odor identification, odor discrimination and olfactory threshold. *Chem Senses* 22:39–52.
27. Bates BT, Zhang S, Dufek JS, Chen FC (1996) The effects of sample size and variability on the correlation coefficient. *Med Sci Sports Exerc* 28:386–391.
28. Kobal G, et al. (1996) "Sniffin' sticks": Screening of olfactory performance. *Rhinology* 34:222–226.
29. Croy I, et al. (2015) Peripheral adaptive filtering in human olfaction? Three studies on prevalence and effects of olfactory training in specific anosmia in more than 1600 participants. *Cortex* 73:180–187.
30. Croy I, et al. (2009) Comparison between odor thresholds for phenyl ethyl alcohol and butanol. *Chem Senses* 34:523–527.
31. Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF (2008) Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biom J* 50:419–430.
32. Doty RL (2006) Olfactory dysfunction and its measurement in the clinic and workplace. *Int Arch Occup Environ Health* 79:268–282.
33. Laska M, Hudson R (1991) A comparison of the detection thresholds of odour mixtures and their components. *Chem Senses* 16:651–662.
34. Oleszkiewicz A, Pellegrino R, Pusch K, Margot C, Hummel T (2017) Chemical complexity of odors increases reliability of olfactory threshold testing. *Sci Rep* 7:39977.
35. Shu CH, Yuan BC, Lin SH, Lin CZ (2007) Cross-cultural application of the "Sniffin' sticks" odor identification test. *Am J Rhinol* 21:570–573.
36. Fornazieri MA, et al. (2015) Development of normative data for the Brazilian adaptation of the University of Pennsylvania Smell Identification Test. *Chem Senses* 40: 141–149.
37. Pierce JD, Jr, Doty RL, Amoore JE (1996) Analysis of position of trial sequence and type of diluent on the detection threshold for phenyl ethyl alcohol using a single staircase method. *Percept Mot Skills* 82:451–458.
38. Doty RL, Gregor TP, Settle RG (1986) Influence of intertrial interval and sniff-bottle volume on phenyl ethyl alcohol odor detection thresholds. *Chem Senses* 11:259–264.
39. Doty RL (2017) Olfactory dysfunction in neurodegenerative diseases: Is there a common pathological substrate? *Lancet Neurol* 16:478–488.
40. Kost RG, Corregano LM, Rainer TL, Melendez C, Coller BS (2015) A data-rich recruitment core to support translational clinical research. *Clin Transl Sci* 8:91–99.
41. Keller A, Vosshall LB (2016) Olfactory perception of chemically diverse molecules. *BMC Neurosci* 17:55.
42. Doty RL (1975) Intranasal trigeminal detection of chemical vapors by humans. *Physiol Behav* 14:855–859.
43. Doty RL, et al. (1978) Intranasal trigeminal stimulation from odorous volatiles: Psychometric responses from anosmic and normal humans. *Physiol Behav* 20:175–185.
44. Kobal G, Van Toller S, Hummel T (1989) Is there directional smelling? *Experientia* 45: 130–132.
45. Croy I, et al. (2014) Human olfactory lateralization requires trigeminal activation. *Neuroimage* 98:289–295.
46. 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.
47. Albrecht J, et al. (2008) Test-retest reliability of the olfactory detection threshold test of the Sniffin' sticks. *Chem Senses* 33:461–467.
48. Deems DA, et al. (1991) Smell and taste disorders, a study of 750 patients from the University of Pennsylvania Smell and Taste Center. *Arch Otolaryngol Head Neck Surg* 117:519–528.
49. Doty RL, Deems DA, Stellar S (1988) Olfactory dysfunction in parkinsonism: A general deficit unrelated to neurologic signs, disease stage, or disease duration. *Neurology* 38:1237–1244.
50. Schwartz BS, Doty RL, Monroe C, Frye R, Barker S (1989) Olfactory function in chemical workers exposed to acrylate and methacrylate vapors. *Am J Public Health* 79:613–618.