Psychometric validity of the Sniffin' Sticks-Extended Test: when to use the sum score confidently and when to be cautious

Eva Tolomeo[1], Carla Masala[2], Antonio Aversa[3], Giancarlo Ottaviano[4], Flavia Gasperi[5,6], Leonardo Menghi[5,6,7], Valentina Parma[8*], Marco Tullio Liuzza[9]

[1] Department of Health Sciences, University "Magna Graecia", Catanzaro, Italy
[2] Department of Biomedical Sciences, University of Cagliari, SP 8 Cittadella Universitaria, 09042 Monserrato, Italy
[3] Department of Experimental and Clinical Medicine, University of Catanzaro "Magna Græcia", 88100 Catanzaro, Italy
[4] Department of Neurosciences, Otolaryngology Section, University of Padova, Padova, Italy
[5] Center Agriculture Food Environment, University of Trento, Via Mach 1, San Michele all'Adige 38098, Italy
[6] Research and Innovation Centre, Fondazione Edmund Mach, Via Mach 1, San Michele all'Adige 38098, Italy
[7] University of Southern Denmark, Department of Technology and Innovation, Campusvej 55, Odense, 5230, Denmark
[8] Monell Chemical Senses Center, Philadelphia, PA
[9] Department of Medical and Surgical Science, University "Magna Graecia", Catanzaro, Italy

***Correspondence:**
Valentina Parma, PhD
Monell Chemical Senses Center
3500 Market Street
Philadelphia, PA 19104
Email: vparma@monell.org

Keywords: Odor detection, discrimination, identification, Sniffin' Sticks, Tau-equivalence, Invariance

Abstract

One commonly used tool to measure olfactory function is the Sniffin' Sticks Test extended version (SSET). The SSET evaluates olfactory ability by summing the scores of three subtests: Threshold, Discrimination, and Identification. Recent meta-scientific literature revealed that many psychometric instruments currently in use have not been adequately validated, leading to a measurement crisis that raises concerns about the validity of the conclusions drawn with these instruments. Two examples of the measurement crisis are i) the use of sum scores without testing their assumptions (i.e., unidimensionality, tau-equivalence, and internal consistency) which indicate that all subtests have the same, stable relationship with their underlying construct, and ii) the lack of assessment of measurement invariance across groups. Here, we aim to investigate the unidimensionality and tau-equivalence assumptions, internal consistency, and measurement invariance of sex and age groups of the Italian version of the SSET. We tested 988 (555 females, mean±SD: 39.75±18.60 years) participants using convenience sampling who were administered the Italian version of the SSET. The tau-equivalent model demonstrated excellent fit indices (CFI robust = 1, TLI robust = 1, RMSEA robust = 0, SRMR = .013), which best explain the data, indicating that that all subtests are equally important in measuring olfactory function, but not necessarily equally precise. The results also revealed full invariance across age groups and configural, partial metric, and scalar invariance across sexes. However, the SSET demonstrated moderate internal consistency. Future studies should clarify whether the reliability of the SSET can be increased, further strengthening its credibility across groups.

Introduction

The COVID-19 pandemic has highlighted like never before the importance of olfactory testing. Millions of people since the beginning of the COVID-19 pandemic have suffered from acute and persistent smell dysfunction (Cecchetto et al., 2021; Doty, 2022; Karamali et al., 2022; Ohla et al., 2022; Tan et al., 2022; Wu et al., 2023), and join additional millions of people who experience olfactory disorders due to different etiologies [e.g., congenital, neurodegenerative disorders, other infectious diseases, metabolic disorders; (Patel et al., 2022)]. Quantifying the prevalence of olfactory disorders is challenging because, at present, there is no system enabling population surveillance of olfactory function (Boesveldt and Parma, 2021). In preparation for this step in healthcare, it is imperative to provide valid measures to evaluate olfactory function across development.

One of the most widely used olfactory tests, especially in Europe, is the Sniffin' Sticks Extended test [SSET; Burghart Messtechnik, Wedel, Germany; (Hummel et al., 1997; Oleszkiewicz et al., 2019)]. In the past 30 years, the SSET has been administered to children (Oleszkiewicz et al., 2019), adults (Wolfensberger, 2000; Neumann et al., 2012; Ribeiro et al., 2016; Niklassen et al., 2018; Oleszkiewicz et al., 2019; Delgado-Losada et al., 2020; Langstaff et al., 2021), and the elderly alike (Oleszkiewicz et al., 2019; Trentin et al., 2022) to assess three olfactory functions: odor threshold (OT) - the ability to detect an odorant concentration 50% of the presentations; odor discrimination (OD) - the ability to distinguish between two different odorants in a triangle test; and odor identification (OI) - the ability to recognize an odorant given four verbal/visual labels. The scores of these three subtests are then summed to produce the TDI score, based on whose cutoffs an individual's olfactory ability is determined (Oleszkiewicz et al., 2019).

The SSET was developed in Germany (Hummel et al., 1997) and it has been translated into several languages, including Italian (Eibenstein et al., 2005; Masala et al., 2022). To our knowledge, none of these translations has been rigorously validated, including assessing the assumptions of using sum scores (i.e., unidimensionality and tau-equivalence), internal consistency, and assessing the measurement validity across age and sex groups, which are variables used to establish norms for olfactory function. As a result, the conclusions reached in the body of work using the SSET might be open to validity issues.

In short, valid measures evaluate what the experimenters assume they are measuring (Kimberlin and Winterstein, 2008). Importantly, measures can not be valid unless they are also reliable, namely accurate and consistent over time (Tavakol and Dennick, 2011). Even if this concept seems trivial, the assumptions underlying whether a measure is valid are paramount to ensuring that science is replicable, and ultimately informative (Lilienfeld and Strother, 2020). As recent meta-scientific research has revealed, most of the measures used in several fields have not been properly validated, raising doubts about the conclusions reached with those tools (Flake et al., 2017; Flake and Fried, 2020).

An essential aspect of validity is the assessment of the structure of a specific construct or **dimensionality** (Pett et al., 2003). A measure or scale is considered unidimensional when there is only one underlying construct or latent variable that accounts for responses to its subtests or items. For instance, the SSET evaluates olfactory function through three subtests. If the SSET is unidimensional, these subtests measure the same construct, which is represented by olfactory function.

There are various models used in confirmatory factor analysis (CFA) to test the structure of a construct based on the underlying assumptions of the measurement (see **Figure S1** for a graphical depiction of the models in supplementary materials). The most restrictive model is the *parallel model* which requires the true loadings of the factors and the error of the measurement to be equal across observations (Raykov, 1997a, 1997b). In simple terms, this means that each subtest or item contributes equally and with the same level of precision to the overall test score, measuring the latent construct. The *tau-equivalent model* requires the true loadings of the factors to remain equal but allows for different error variances among observations. This implies that each subtest is equally related to the underlying construct, but the precision of each subtest can vary, meaning that all subtests are equally important in measuring the construct, but not necessarily equally precise. The least restrictive model is the *congeneric model*, which allows both factor loadings and error variances to vary. This means that subtests can differ in their degree of contribution and precision when measuring the underlying construct, acknowledging that some subtests may be better indicators of the construct than others. Oftentimes, a congeneric model is assumed even though the measure (e.g., a sum score) requires the assumptions of a unidimensional or tau-equivalent model to be met (McNeish and Wolf, 2020).

To determine whether the TDI score is the appropriate summary measure of olfactory function when using the SSET, there are two possible options: testing for unidimensionality or tau-equivalence. McNeish and Wolf (McNeish and Wolf, 2020) claim that the valid use of a sum score requires unidimensionality and a parallel model, though others argue that unidimensionality is a sufficient condition (Widaman and Revelle, 2022),

supporting the suitability of tau-equivalent models. Tau equivalence indicates that each item or subtest exhibits equal factor loadings, which are coefficients that represent the strength of the relationship between each subtest and the underlying construct being measured - in this case, olfactory function. Equal factor loadings across the three subtests imply that each subtest contributes equally to the overall assessment of the construct. Therefore, the scores from each subtest are equally significant and informative when combined to produce a comprehensive measure of olfactory function. Meeting the tau equivalence assumption allows us to aggregate the scores of OT, OD, and OI to derive the TDI score, ensuring its validity as a measure. A further step would include evaluating whether having a deficit in one subtest could undermine the validity of TDI in predicting olfactory function. To our knowledge the underlying model of the SSET has not been tested, leaving its validity open to failure.

To assess the reliability of the TDI score, we need to measure its internal consistency. This involves determining if the three subtests of the SSET measure the same construct or latent variable. Acceptable internal consistency indices (such as Cronbach's $\alpha > .7$) establish an upper limit for validity, given that an instrument cannot be more valid than reliable (Tavakol and Dennick, 2011). However, it is important to note that the value of Cronbach's α is affected by the number of items. In general, Cronbach's α tends to be lower in tests with fewer items, since it is a function of both the number of items and the correlations among them. In the case of the SSET, Cronbach's α is determined based on the OT, OD, and OI scores, using only three items. To achieve a Cronbach's α greater than .7 with just three items, the average correlation across all items should be at least 0.43.

Additionally, the SSET is often used to compare the olfactory ability of different groups of individuals, such as males and females, or the same individuals over time. These comparisons highlight the importance of testing another crucial psychometric property - **measurement invariance**, which assesses that the measured construct is psychometrically equivalent across groups and/or times. To our knowledge, measurement invariance has never been tested for the SSET, nor for the other olfactory measures currently used, with the consequence that all results based on groups and/or time comparisons could be misleading. Meeting the criteria for measurement invariance indicates that the construct has the same structure across different groups or time points, and this allows meaningful within-group and between-group comparisons (Putnick and Bornstein, 2016).

Measurement invariance is usually tested through three main steps (Widaman and Reise, 1997). First, one should perform a *configural invariance test*. This test allows us to estimate factor loadings, intercepts, and residuals freely, to establish a baseline model across groups. The configural invariance test implies that similar, even though not identical, latent constructs have been measured in the groups (Widaman and Reise, 1997). If the configural invariance test is not supported by the data, the measurement invariance test should not proceed further. Second, a *metric invariance model* is tested, constraining all the factor loadings to be equal. This is a weak invariance test, and it is considered successful when the relationship between indicators and latent variable(s) is invariant across groups. In other words, weak invariance holds when a unit increase in the latent variable is reflected by the same increase in the observed variable across groups. Third, a *scalar invariance model* is tested. In this model, the factor loadings and

the intercepts are constrained to be the same across groups: this is considered a strong invariance test, and it is required to compare latent means across groups. Indeed, if scalar invariance is not met, any group difference in the observed scores may not reflect a real difference in the latent means or, vice versa, an apparent lack of difference in the observed scores may obscure a real difference in the latent means.

Based on these premises, the overarching aim of the present work is to psychometrically validate the SSET in the Italian population by assessing its unidimensionality, tau-equivalence, internal consistency, and measurement invariance. First, we assessed the unidimensionality of the SSET to ensure that all three subtests are measuring the same construct. If all subtests of the SSET highly correlate with each other, in line with Widaman and Revelle (2022), and confirmatory factor analysis reveals good fit, we consider the unidimensionality assumption met. Second, we assessed if the SSET is based at least on a tau-equivalent model, to justify the use of the TDI sum score as a summary measure. To determine this, we examined the fit of the tau-equivalent model and its explanatory power compared to the parallel and congeneric models. If the tau-equivalent model demonstrates good fit measures, and better fit than the other models, the tau-equivalence assumption is confirmed and the TDI score is considered a valid summary measure for the SSET. Third, we assessed whether deficits in one subtest could affect the validity of the TDI sum score. Fourth, we explored whether the internal consistency of the SSET is acceptable. Finally, we evaluated measurement invariance across age and sex groups so that the test scores could be used to make meaningful comparisons across groups and time. If full invariance is achieved, then the observed means (i.e., the raw TDI score) can directly be used to compare groups. If only partial

invariance is supported, the latent means (i.e., means estimated through factor analysis) should be compared.

Materials & Methods

Participants

We recruited 999 participants using convenience sampling who were administered the Italian version of the SSET. We excluded 11 of them because we did not share their sex (n = 1), and age (n = 8), or were missing the score at the OT subtest (n = 2; **Table 1**). The final sample consisted of 988 subjects [555 F (56%), mean±SD of age: 39.75±18.60 years, age range: 18-92 years] from five different Italian regions [Calabria (n = 187), Friuli-Venezia Giulia (n = 30), Sardinia (n = 466), Trentino (n = 89), and Veneto (n = 216)]. To test measurement invariance across ages, participants were divided into three age groups: 18-30 (n = 483), 31-60 (n = 322), and 61+ (n = 183). Sex was coded as a binary variable (F/M) given that no other option was reported. Data was collected in the context of different studies (Iuliano et al., 2023; Masala et al., 2022), and collapsed in this analysis. The exclusion criteria used across participants were: 1) history of neurologic or psychiatric disease, 2) head injury with loss of consciousness, 3) local respiratory tract diseases at the time of the testing (such as allergic or infectious rhinitis, or sinusitis), 4) pregnancy, and 5) any ongoing oncological treatment. Eligibility criteria were checked by the examiner before the beginning of the testing.

Table 1. Descriptive statistics of age and SSET subtests.

	SSET

	Age	OT	OD	OI
N	988	988	988	988
Mean	39.75	7.74	12.02	12.73
SD	18.6	3.79	2.19	2.02
Minimum	18	1	1	1
Maximum	92	16	16	16
Skewness	0.8	0.46	-0.93	-1.23
Kurtosis	-0.62	-0.32	1.91	3.06

Note. SSET: Sniffin' Sticks Extended test. OT: odor threshold subtest. OD: odor discrimination subtest. OI: odor identification subtest. SD: standard deviation. N = 156 (16%) participants show at least one deficit in one subtest. Please see Table S1 for further detail on deficits across subtests.

The Sniffin' Sticks Extended Test

The SSET is composed of three subtests: OT, OD, and OI. From each subtest, we obtain a score ranging from 0 to 16, and the sum of the scores of the three subtests represents the composite TDI score, which ranges from 0 to 48. The TDI score is commonly interpreted as a measure of olfactory function (Hummel et al., 1997). The examiner administers the test wearing odorless gloves. Each odorous pen's tip is placed about 2 cm from the subject's nostrils. Each pen is presented for about 3 seconds.

The OT subtest measures the lowest concentration of odorant that an individual can detect in 50% of presentations. It comprises 16 triplets of odorous pens and in each triplet, only one pen contains n-butanol, while the other two pens contain a distractor solvent. In each trial, three pens are presented in a randomized order, and the participant's task is to indicate the pen containing the odorant (n-butanol) among the two that do not. The interval between presentations of each triplet is approximately 20s. The concentration of the odorant varies across triplets. The examiner presents a triplet with an ascending odor concentration, starting from the lowest concentration. After two

consecutive correct responses, a triplet with a descending odor concentration is presented. The OT score is calculated as the mean of the last four reversals and ranges from 16 – indicating that participants can detect the lowest concentration of the odorant – to 1 – indicating participants are unable to detect the highest concentration of the odorant.

The OD subtest evaluates the ability of a person to discriminate if two suprathreshold odorants are different or not. The test is composed of 16 triplets of pens. Each triplet contains two pens filled with the same non-target odor, and one pen with a different target odor. In each trial, the participant's task is to identify the target pen. For each correct answer, one point is assigned. The score ranges from 0 to 16 and it is the sum of correct answers. The higher the score, the better the OD ability.

The OI test evaluates the ability of a participant to correctly recognize an odorant at a suprathreshold level among four given verbal and visual options. The test consists of 16 pens containing different odorants commonly encountered in everyday life (i.e., flower, lemon, fish). The examiner presents one pen at a time and the participant’s task is to select the accurate response from the four options provided. The score ranges from 0 to 16 and it is the sum of correct answers.

Procedures

Olfactory function was assessed with the SSET. The use of the test was approved by the Ethics Committees of the University Hospital of Cagliari (Prot. Number: NP/2018/1630) and the Internal Review Board of A.O.U. Renato Dulbecco, Catanzaro, and the procedures were carried out according to the Declaration of Helsinki. All participants gave their written informed consent to participate in the study. Participants were tested

individually and completed three subtests of the SSET in their established order (OT, OD, OI). Across studies, participants were asked not to eat and drink only water for at least 1 h before the testing session. The procedure lasted about 40-60 minutes and was carried out in a quiet and well-ventilated room.

Data Analyses

Statistical analyses were performed with R (R Core Team, 2023). We assessed multivariate normality using Mardia's test [assumption met for $p > 0.05$; (Mardia, 1970)], revealing that multivariate normality was not supported [Mardia Skewness = 424.64 ($p < .001$), Mardia Kurtosis = 14.05 ($p < .001$)]. In the absence of a normal distribution, we employed a Maximum Likelihood Robust (MLR) estimator for the confirmatory factor analyses (CFA). Then, we investigated whether the assumptions required for summing the scores of the three subtests - unidimensionality and tau-equivalence - were met. To assess the unidimensionality of the SSET we performed a confirmatory factor analysis (CFA) using the R package *lavaan* (Rosseel, 2012). The goodness of the model fit was evaluated using the following indices: Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), Tucker–Lewis Index (TLI), and Standardized Root Mean Square Residual (SRMR). Values below 0.05 indicated a good fit for RMSEA and SRMR, while for CFI and TLI values above 0.90 were considered acceptable (Hu and Bentler, 1999).

Whenever evidence of tau-equivalence is revealed (McNeish and Wolf, 2020), internal consistency can be assessed using Cronbach's α (Cronbach, 1951). Values of α higher than .70 were considered acceptable, even if there is limited evidence that such

value should be considered a decisional threshold (Taber, 2018), especially with a limited number of items (n = 3), as in our case.

In absence of large-scale independent norms for the Italian population, participants were considered as having deficits if their scores fell below the tenth percentile in OT (<3.25), OD (<9), or OI subtests (<10, see Table S1). To investigate whether a deficit in one subtest influences the relationship between the other two subtests, we conducted a set of moderated regression analyses using three models. Each model addressed whether the relationship between each pair of subtests was moderated by a deficit in the other subtest. We then followed-up on the results by checking whether the relationship between each pair was significantly different from zero in each subgroup (deficit and control group). We chose this approach rather than a measurement invariance approach due to the lack of statistical power for the latter.

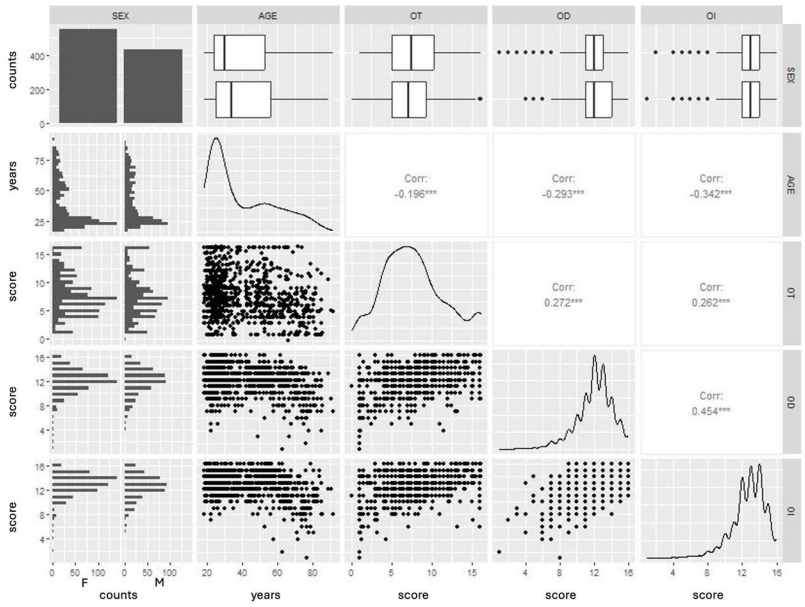
Measurement invariance across sexes and ages groups was tested with progressively restrictive criteria, according to Vandenberg and Lance (2000). We tested measurement invariance using the congeneric model, and we fixed the variance of the latent variable to 1. We started testing a configural invariance model, then a metric model, and a scalar model. Measurement invariance is rejected when the comparison between two models leads to a significant $\Delta\chi 2$ ($p < .05$). However, since this method is prone to easily reject measurement invariance (Hays et al., 2005), measurement invariance was assessed by considering the differences in CFI, RMSEA, and SRMR. In detail, changes ≤ - 0.010 in ΔCFI, ≥ 0.015 in ΔRMSEA or ≥ 0.030 in ΔSRMR indicate non-invariance for metric invariance, while changes ≤ - 0.010 in ΔCFI, ≥ 0.015 in ΔRMSEA, or ≥ 0.015 in ΔSRMR indicate non-invariance for scalar invariance (Chen, 2007).

Results

The latent variable of the SSET is unidimensional

To determine whether the TDI score of the SSET reflects a unidimensional construct (i.e., olfactory function), we first assessed the correlations across the three subtests of the SSET (**Figure 1**). OT showed a weak correlation with OD (*r* =.27) and OI (*r* = .26). In contrast, a stronger correlation was detected between OI and OD (*r* = .45). Furthermore, results of the confirmatory factor analysis highlighted that the tau-equivalent model showed excellent fit indices (CFI robust = 1, TLI robust = 1, RMSEA robust = 0, SRMR = .013), thus confirming the unidimensional structure of the SSET.

Figure 1. Frequency, boxplots, density, and correlations for sex, age, OT, OD, and OI.



Note. F: females. M: males. OT: odor threshold subtest. OD: odor discrimination subtest. OI: odor identification subtest. Corr: Pearson's correlation. * $p < .05$; ** $p < .01$; *** $p < .001$.

Using the TDI score is justified by the satisfaction of the tau-equivalence assumption

To explore whether the use of the TDI as a sum score is justified, we fitted three measurement models (i.e., parallel, tau-equivalent, and congeneric). Results of the CFA revealed that the tau-equivalent model showed excellent fit indices (CFI robust = 1, TLI

robust = 1, RMSEA robust = 0, SRMR = .013), and the fit did not significantly worsen when compared to the unidimensional congeneric model ($\Delta\chi 2$ = 1.26, *df* = 2, *p* =.534, **Table 2**). An analysis of AIC and BIC across the three models revealed that the tau-equivalent model was the model that best fit the data (**Table 2**), revealing that the tau-equivalence assumption in the Italian version of the SSET can be assumed.

Table 2. Comparison between parallel, tau-equivalent, and congeneric models.

Model	AIC	BIC	$\chi 2$	df	$\Delta\chi 2$	Δdf	*p*
Congeneric	13662	13692	0.00	0			
Tau-equivalent	13660	13679	1.18	2	1.25	2	.535
Parallel	14288	14298	634.07	4	378.22	2	< .001

Note: AIC: Akaike Information Criterion; BIC: Bayesian Information Criterion; $\chi 2$: chi-squared test; df: degrees of freedom; $\Delta\chi 2$: chi-squared test difference; Δdf: degrees of freedom difference; *p* = p-value.

The validity of the SSET persists even for participants with olfactory deficits

Moderated regression analyses on participants that showed deficits in each subtest were employed to further explore the interrelationships among the subtests. Results reported in Table 3 indicate that a deficit in OI significantly impacts the OD-OT relationship, as revealed by significant main effects for OT and deficitOI on OD and a significant interaction term. Additionally, results indicate that a deficit in OD does not significantly alter the OT-OI relationship even though the main effects for OT and deficitOD on OI were significant. Finally, results indicate that a deficit in OT significantly impacts the OI-OD relationship (significant main effects for OD and deficitOI on OT and significant interaction term). Moreover, across groups, we consistently observed high and statistically significant correlations: in the OT deficit group (n = 97), OD correlated with OI r = .72 (p < .001); in the OD deficit group (n = 64), OT correlated with OI r = .46 (p < .001); and in the OI deficit group (n = 62), OT correlated with OD r = .58 (p < .001).

Table 3. Output of moderate regression analysis on groups of participants with olfactory deficits.

Model	Effect	Estimate	Standard Error	t value	p-value
OD ~ OT * deficitOI	Intercept	11.450	0.150	76.388	< .001
	OT	0.094	0.017	5.463	< .001
	deficitOI	-4.915	0.426	-11.532	< .001
	OT:deficitOI	0.337	0.068	4.977	< .01
OT ~ OI * deficitOD	Intercept	3.381	0.899	3.761	< .001
	OI	0.354	0.069	5.151	< .001

	deficitOD	-3.88	1.747	-2.219	< .01
	OI:deficitOI	0.158	0.163	0.972	0.331
OI ~ OD * deficitOT	Intercept	9.524	0.372	25.549	< .001
	OD	0.278	0.03	9.216	< .001
	deficitOT	-5.717	0.699	-8.182	< .001
	OD:deficitOT	0.449	0.064	7.013	< .001

Internal consistency of the SSET is low, due to high variance in OT

Given the acceptance of the tau-equivalent model, we estimated internal consistency across the three subtests of the SSET using Cronbach's α. Results revealed that the estimate did not reach acceptable levels (α = .53), suggesting large amounts of measurement error. Estimates of residual variance (OT = 12.24, OD = 2.73, OI = 2.09) indicated that a relevant amount of measurement error is derived from the OT subtest. Moreover, the low number of items likely contributed to the lower α value. Furthermore, the average inter-item (i.e., item = SSET subtest) correlation was .33, lower than the needed .43, necessary to achieve Cronbach's α of .7. In fact, OD and OI have a correlation of *r* = .45, but OT showed weaker correlation with the other subtests (OT-OD *r* = .27; OT-OI *r* = .26). While OD and OI subtests share .47 and .44 variance with the latent variable, OT only shares .16 with the latent variable.

Measurement invariance across age groups is met

To investigate measurement invariance across age, we conducted a multiple-group CFA on the three age groups (18-30, 31-60, and 61+ years old). Results showed that full configural, metric, and scalar invariance were met (**Table 4**).

Table 4. Fit indices for invariance testing across age groups.

	AIC	BIC	$\chi 2$	df	$\Delta\chi 2$	Δdf	p
Configural invariance	13392	13524	0	0			
Metric invariance	13390	13502	5.41	4	6.08	4	.193
Scalar invariance	13385	13478	8.76	8	3.16	4	.530

Note: AIC: Akaike Information Criterion; BIC: Bayesian Information Criterion; $\chi 2$: chi-squared test; df: degrees of freedom; $\Delta\chi 2$: chi-squared test difference; Δdf: degrees of freedom difference; p = p-value.

First, we conducted an unconstrained analysis of the three age groups to confirm the configural invariance of the SSET. Second, we constrained the factor loadings to be equal between the three age groups for the metric invariance analysis. Comparison between the configural and metric the model suggested that the metric model did not worsen the fit, so metric invariance was achieved ($\Delta\chi 2$ = 5.06, p = .192). Third, we performed the scalar invariance analysis by constraining the factor loadings and intercepts of the items to be equal across age groups. Results showed that the scalar model fit did not worsen when compared to the metric model, indicating that scalar invariance was met ($\Delta\chi 2$ = 0.39, p = .531).

Since full invariance was met, we proceeded by comparing the observed means of the three age groups performing a one-way ANOVA. The ANOVA revealed that there was a statistically significant difference in olfactory function (F(2, 985) = 77.96, p < .001). Tukey's HSD Test for multiple comparisons found that the mean value of olfactory function was significantly lower for the older age group than the younger age group (p < .001, 95% C.I. = [-6.98, -4.71]), and the median age group (p < .001, 95% C.I. = [-6.62, -4.20]). On the other hand, there was no statistically significant difference between the younger age group and the median age group (p = .523). The effect size, calculated as eta squared (η^2), was 0.14, indicating a medium effect.

Measurement of invariance across sexes is partially met

We also tested the SSET for measurement invariance across sexes. We started testing configural invariance performing an unconstrained analysis. Then, we tested metric invariance constraining the factor loadings to be equal between males and females. Results of the comparison between the configural and the metric model revealed that the metric model worsened the fit, suggesting that there is a lack of metric invariance for the SST across sexes ($\Delta\chi 2 = 15.26$, $p > .001$). For this reason, we tried to establish partial invariance. We proceeded to explore the parameters in order to identify which fixed (or constrained) parameters in the model should be released to improve the fit for the scalar model. We found that the OD loading varied across males and females, and after relaxing this parameter, partial metric invariance was met. Then, we proceeded by testing for full scalar invariance by constraining the factor loadings and intercepts of the items to be equal across sex groups. The scalar model did not worsen the fit when compared to the partial metric model, thus indicating that scalar invariance was met. The results of fit indices for invariance testing are reported in **Table 5.**

Table 5. Fit indices for invariance testing across sex

	AIC	**BIC**	$\chi 2$	**df**	$\Delta\chi 2$	Δdf	p
Configural invariance	13631	13719	0	0			
Partial Metric invariance	13629	13712	0.13	1	0.15	1	0.699
Scalar invariance	13630	13704	5.38	3	5.26	2	0.072

Note: AIC: Akaike Information Criterion; BIC: Bayesian Information Criterion; $\chi 2$: chi-squared test; df: degrees of freedom; $\Delta\chi 2$: chi-squared test difference; Δdf: degrees of freedom difference; p = p-value.

We compared the latent means of males and females and found that they significantly differ (Estimate = -0.25, SE = 0.10, z = -2.36, p = .018). We also compared the observed means. First, Levene's test was conducted to assess the homogeneity of variances across groups. The result was significant [F(1,986) = 17.81, p < .001], indicating that the assumption of equal variances was violated. For this reason, we performed a Welch's t-test, which is robust to unequal variances. Results highlighted that they were statistically different [t(818.77) = 2.4, p = .016], with females obtaining higher scores (M = 32.90) than men (M = 31.96). The effect size, as measured by Cohen's d, was d = 0.16, indicating a small effect. To calculate observed and latent means by sex for the SSET, please refer to the following R script available on OSF: https://osf.io/djpae/.

Discussion

The main goal of this study was to validate the TDI score in the Italian version of the SSET as a valid and reliable measure of olfactory function, for which the unidimensionality, tau-equivalence assumptions, internal consistency, and measurement invariance are tested.

The results from the confirmatory factor analysis supported the unidimensional structure of the SSET. Moreover, the tau-equivalent model (same loadings, different error variances) showed excellent fit indices and, when compared to the congeneric model (different loadings and different error variances), it did not significantly worsen the fit, indicating that it provides an adequate explanation of the data. These findings suggest that the assumptions of unidimensionality and tau-equivalence are met for the SSET in the Italian population, allowing the use of the sum score (TDI) of three subtests as a valid measure of olfactory function.

Moreover, the validity of the TDI is not affected by having a deficit in one subtests. Both correlations and moderated regression analyses suggest that despite deficits in each individual subtest, associations among the other subtest scores are significantly different from zero. This indicates that the SSET comprehensively assesses olfactory function, even in presence of specific olfactory dysfunction. These results provide substantial evidence affirming the validity of combining the three measures in the SSET, thus confirming its utility in accurately assessing olfactory function.

With respect to reliability, the internal consistency estimate of the three subtests (α = .53) did not reach acceptable levels, thus revealing the presence of a large amount of measurement error. Residual variance estimates revealed significant measurement errors originating from the OT subtest. This variability arises partly because the other two subtests (OD and OI) exhibit a ceiling effect (future studies should aim to make these subtests more challenging in the Italian version) and partly because OT is more susceptible to error (due to testing conditions, etc.). Future research efforts should focus on addressing these factors to improve the robustness of the SSET to reach acceptable Cronbach's α values. It is important to note that factors such as the number of test items can influence the value of α (Cortina, 1993): the lower the number of items (here subtests), the lower the internal consistency. Additionally, results demonstrated a moderate average inter-item correlation, suggesting satisfactory homogeneity of the items, but also a notable amount of unique variance. In other words this means that the OT, OD and OI subtest partially overlap in explaining olfactory function, but not completely. It is worth mentioning that the subtests of the TDI score include metrics of odor sensitivity, discrimination, and identification providing a multi-pronged representation of olfactory function, leaving the possibility of further improvement in internal consistency uncertain.

An alternative account would be to conceive the TDI score as a formative model instead of a reflective one (Avila et al., 2015). This means that rather than thinking of

olfactory function as a latent variable that causes a good performance in the OT, OD, and OI, we may conceive the performance at these subtests as causing the latent variable. Under this model, correlation among the formative indicators is not required, and the use of Cronbach's α as an estimate of internal consistency would be unwarranted.

The analysis of measurement invariance across ages revealed that the SSET in the Italian population measures the same construct (i.e., olfactory function) for all age groups since full configural invariance was met. In other words, this means that the factor is associated with the same set of items across groups (Gregorich, 2006). The SSET displayed also metric invariance, which provides evidence that the factor, and its relationship with the responses to a common set of items, has the same meaning across the three age groups. Furthermore, results highlighted that the SSET is scalar invariant across age groups, meaning that mean differences in the latent factor catch all mean differences in the shared variance of the item (Putnick and Bornstein, 2016). At this level, the interpretation of group comparisons is not biased by the presence of confounding variables. Since the SSET showed full invariance, observed means of groups can be compared across age groups. From the mean comparison of the three age groups, we found that the median group did not have significantly lower levels of olfactory function than the younger group. Conversely, the elderly group had significantly lower levels than the younger group and the median group. This result is in line with evidence demonstrating that olfactory function declines with aging (Doty and Kamath, 2014; Masala et al., 2018; Hummel and Oleszkiewicz, 2020). In particular, the incidence of the dysfunction increases from age 60 onwards (Doty et al., 1984; Shiffman, 1997; Kondo et al., 2020), and this explains why the younger group did not differ from the median group.

Analysis of measurement invariance across sexes showed evidence for configural invariance, demonstrating that the SSET is measuring the same construct among males and females. Yet, metric invariance was only partially met, given that males displayed higher loadings compared to females in the OD subtest. However, the SSET was scalar

invariant, indicating that the intercepts, or baseline levels, of olfactory ability are the same for both male and female groups. Partial invariance allows meaningful comparison of latent means across males and females, and results indicated that they significantly differ in olfactory function. In detail, females reported higher scores than males. The same result emerged from the comparison of observed means, although the effect size was very small. In this case, the comparison of observed means is not particularly problematic, but less reliable than the comparison of latent means. Full measurement invariance is often not met in all steps, and this leads to accepting as a common practice some violations of measurement invariance (such as the release of constraints on loadings and/or intercepts) and carrying on analyses using the partial invariant factor (Putnick and Bornstein, 2016). Steinmetz (2013) performed a Monte Carlo simulation study to investigate whether unequal factor loadings and intercepts affect conclusions on latent mean differences and found that metric non-invariance is less problematic while scalar non-invariance can lead to weighty misinterpretation of true mean differences. This result is important because it confirms that when looking at observed scores (i.e. the raw TDI data), our results align with the literature suggesting females' SSET scores were greater than those of men (Brand and Millot, 2001; Oleszkiewicz et al., 2019; Masala et al., 2022). For instance, a recent meta-analysis by Sorokowski et al. (2019) indicates that when considering observed measures, women perform better than men in all aspects of olfactory function tested using the SSET, including OT, OD, and OI. However, these effect sizes are small, revealing that sex differences account for less than 1% of the variability in the SSET. Furthermore, there is a significant variation in effect sizes when comparing the OI subtest of the SSET (Hedge's g = 0.08) with another commonly used odor identification test, the University of Pennsylvania Smell Identification Test (UPSIT) [Hedge's g = 0.3, (Doty et al., 1984b)]. These findings suggest that sex differences in olfactory function might be largely influenced by issues of measurement invariance since similar tests measuring the same function (UPSIT and SSET's OI) should yield similar

sex differences. The availability of datasets with olfactory test scores and sex information will help settle this controversy by allowing for the direct testing of observed and latent scores.

In summary, the analysis of measurement invariance suggests that the sum score of the SSET can be reliably used to compare observed means across age groups, since full invariance was met. With regard to sexes, partial invariance permits meaningful comparisons of latent means between males and females. Comparisons of observed means are also possible, but less reliable.In the end, the SSET proved to be a valid measure of olfactory function. The moderate reliability could be addressed by improving the subtests of the Italian version. In this regard, further analyses of the data using an Item Response Theory approach (Embretson and Reise, 2013) would help to evaluate whether the items used in the OD and OI have the same discrimination power and whether they cover a sufficiently large range of abilities. Indeed, conducting a so-called differential item functioning (DIF) analysis based on sex and age can uncover any potential biases inherent in the items.

Limitations

Our study has some limitations. First, the sample tested may not be representative of the entire Italian population, given that the sampling strategy used for this study was one of convenience. Second, we used a latent variable model perspective, which assumes causal relations between measures, and the latent variable is considered as the underlying factor that explains the observed variables.

Future studies using a network analysis approach could open opportunities to address the weaknesses revealed here concerning internal consistency. According to this approach, psychological attributes are considered complex systems in which each component is in interaction with and influenced by the others, but without a common cause that explains them (Guyon et al., 2017; Borsboom et al., 2021). For instance, rather

than conceiving OT, OD, and OI as caused by a latent olfactory function, we may investigate how these performances may affect each other, and identify which of them plays a central role in the network.

Conclusions

For the first time, we investigated whether the use of the TDI score of the SSET in its Italian version is a valid and reliable measure of olfactory function. Unidimensionality and tau-equivalence properties are satisfied, enhancing the validity of the SSET. However, internal consistency was moderate. Moreover, the TDI score should not be used to compare observed means of sex groups directly, but rather latent means should be used. Future research is crucial to determine if the SSET reliability can be enhanced based on Item Response Theory insights, thereby bolstering the credibility of the SSET findings among different groups.

Competing interests

The authors declare no conflict of interest.

Author contributions

ET: data analysis, manuscript writing and reviewing; CM: conceptualization, data collection, reviewing; AA: data collection, manuscript reviewing; GO: data collection, manuscript reviewing; FG: data collection, manuscript reviewing; LM: data collection, manuscript reviewing; MPC: data collection, manuscript reviewing; VP: conceptualization, data collection, manuscript writing, and reviewing; MTL: conceptualization, data collection, data analysis, manuscript writing, and reviewing.

Funding

The current research was funded by MIUR-PRIN (Ministry of University and Research-Projects of National Relevant Interest) granted to M.T.L. (grant number: 2022TN4ETY).

Acknowledgments

This work would not be possible without people who are willing to participate in the research.

Data Availability

The datasets and scripts generated and/or analyzed during the current study are available in the OSF repository upon acceptance of the publication, https://osf.io/djpae/. In case unforeseen issues will prevent the access to the OSF repository, the datasets used and/or analyzed during the current study are available from the corresponding author upon request.

References

Avila, M.L., Stinson, J., Kiss, A., Brandão, L.R., Uleryk, E., and Feldman, B.M. 2015. A critical review of scoring options for clinical measurement tools. BMC Res Notes. 8:612.

Boesveldt, S., and Parma, V. 2021. The importance of the olfactory system in human well-being, through nutrition and social behavior. Cell Tissue Res. 383:559–567.

Borsboom, D., Deserno, M.K., Rhemtulla, M., Epskamp, S., Fried, E.I., McNally, R.J., Robinaugh, D.J., Perugini, M., Dalege, J., Costantini, G., et al. 2021. Network analysis of multivariate data in psychological science. Nat Rev Methods Primers. 1:58.

Brand, G., and Millot, J.-L. 2001. Sex differences in human olfaction: Between evidence and enigma. The Quarterly Journal of Experimental Psychology B. 54:259–270.

Cecchetto, C., Di Pizio, A., Genovese, F., Calcinoni, O., Macchi, A., Dunkel, A., Ohla, K., Spinelli, S., Farruggia, M.C., Joseph, P.V., et al. 2021. Assessing the extent and timing of chemosensory impairments during COVID-19 pandemic. Sci Rep. 11:17504.

Chen, F.F. 2007. Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. Structural Equation Modeling: A Multidisciplinary Journal. 14:464–504.

Cortina, J.M. 1993. What is coefficient alpha? An examination of theory and applications. Journal of Applied Psychology. 78:98–104.

Cronbach, L.J. 1951. Coefficient alpha and the internal structure of tests. Psychometrika. 16:297–334.

Delgado-Losada, M.L., Delgado-Lima, A.H., and Bouhaben, J. 2020. Spanish Validation for Olfactory Function Testing Using the Sniffin' Sticks Olfactory Test: Threshold, Discrimination, and Identification. Brain Sciences. 10:943.

Doty, R.L. 2022. Olfactory dysfunction in COVID-19: pathology and long-term implications for brain health. Trends in Molecular Medicine. 28:781–794.

Doty, R.L., and Kamath, V. 2014. The influences of age on olfaction: a review. Front Psychol. 5.

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

Doty, R.L., Shaman, P., and Dann, M. 1984b. Development of the university of pennsylvania smell identification test: A standardized microencapsulated test of olfactory function. Physiology & Behavior. 32:489–502.

Eibenstein, A., Fioretti, A.B., Lena, C., Rosati, N., Ottaviano, I., and Fusetti, M. 2005. Olfactory screening test: experience in 102 Italian subjects. Acta Otorhinolaryngol Ital. 25:18–22.

Embretson, S.E., and Reise, S.P. 2013. Item Response Theory. Psychology Press.

Flake, J.K., and Fried, E.I. 2020. Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. Advances in Methods and Practices in Psychological Science. 3:456–465.

Flake, J.K., Pek, J., and Hehman, E. 2017. Construct Validation in Social and Personality Research: Current Practice and Recommendations. Social Psychological and Personality Science. 8:370–378.

Gregorich, S.E. 2006. Do Self-Report Instruments Allow Meaningful Comparisons Across Diverse Population Groups?: Testing Measurement Invariance Using the Confirmatory Factor Analysis Framework. Medical Care. 44:S78.

Guyon, H., Falissard, B., and Kop, J.-L. 2017. Modeling Psychological Attributes in Psychology - An Epistemological Discussion: Network Analysis vs. Latent Variables. Front Psychol. 8:798.

Hays, R.D., Revicki, D., and Coyne, K.S. 2005. Application of structural equation modeling to health outcomes research. Eval Health Prof. 28:295–309.

Hu, L., and Bentler, P.M. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal. 6:1–55.

Hummel, T., and Oleszkiewicz, A. 2020. Age-Related Changes of the Sense of Smell. In: The Senses: A Comprehensive Reference. Elsevier. pp. 717–726.

Hummel, T., Sekinger, B., Wolf, S.R., Pauli, E., and Kobal, G. 1997. 'Sniffin' Sticks': Olfactory Performance Assessed by the Combined Testing of Odour Identification, Odor Discrimination and Olfactory Threshold. Chem Senses. 22:39–52.

Iuliano, S., Liuzza, Prof.M.T., Zagari, Dr.M.C., Vergine, Dr.M., Seminara, Dr.G., Vaccaro, Dr.M.G., Tolomeo, E., Sgro, A., Rizzuto, F., Carlo, Prof.C.D., et al. 2022. Smells like good sex: olfactory function is associated with sexual well-being among fertile women. The Journal of Sexual Medicine. 19:S111.

Karamali, K., Elliott, M., and Hopkins, C. 2022. COVID-19 related olfactory dysfunction. Current Opinion in Otolaryngology & Head & Neck Surgery. 30:19–25.

Kimberlin, C.L., and Winterstein, A.G. 2008. Validity and reliability of measurement instruments used in research. American Journal of Health-System Pharmacy. 65:2276–2284.

Kondo, K., Kikuta, S., Ueha, R., Suzukawa, K., and Yamasoba, T. 2020. Age-Related Olfactory Dysfunction: Epidemiology, Pathophysiology, and Clinical Management. Front Aging Neurosci. 12:208.

Langstaff, L., Clark, A., Salam, M., and Philpott, C.M. 2021. Cultural Adaptation and Validity of the Sniffin' Sticks Psychophysical Test for the UK Setting. Chem Percept. 14:102–108.

Lilienfeld, S.O., and Strother, A.N. 2020. Psychological measurement and the replication crisis: Four sacred cows. Canadian Psychology / Psychologie Canadienne. 61:281–288.

Mardia, K.V. 1970. Measures of multivariate skewness and kurtosis with applications. Biometrika. 57:519–530.

Masala, C., Cavazzana, A., Sanna, F., Cecchini, M.P., Zanini, A., Gasperi, F., Menghi, L., Endrizzi, I., Borgogno, M., Drago, S., et al. 2022. Correlation between olfactory function, age, sex, and cognitive reserve index in the Italian population. Eur Arch Otorhinolaryngol. 279:4943–4952.

Masala, C., Saba, L., Cecchini, M.P., Solla, P., and Loy, F. 2018. Olfactory Function and Age: a Sniffin' Sticks Extended Test Study Performed in Sardinia. Chem Percept. 11:19–26.

McNeish, D., and Wolf, M.G. 2020. Thinking twice about sum scores. Behav Res. 52:2287–2305.

Neumann, C., Tsioulos, K., Merkonidis, C., Salam, M., Clark, A., and Philpott, C. 2012. Validation study of the "Sniffin' Sticks" olfactory test in a British population: a preliminary communication. Clinical Otolaryngology. 37:23–27.

Niklassen, A.S., Ovesen, T., Fernandes, H., and Fjaeldstad, A.W. 2018. Danish validation of sniffin' sticks olfactory test for threshold, discrimination, and identification. The Laryngoscope. 128:1759–1766.

Ohla, K., Veldhuizen, M.G., Green, T., Hannum, M.E., Bakke, A.J., Moein, S.T., Tognetti, A., Postma, E.M., Pellegrino, R., Hwang, D.L.D., et al. 2022. A follow-up on quantitative and qualitative olfactory dysfunction and other symptoms in patients recovering from COVID-19 smell loss. Rhinology Journal. 0.

Oleszkiewicz, A., Schriever, V.A., Croy, I., Hähner, A., and Hummel, T. 2019. Updated Sniffin' Sticks normative data based on an extended sample of 9139 subjects. Eur Arch Otorhinolaryngol. 276:719–728.

Patel, Z.M., Holbrook, E.H., Turner, J.H., Adappa, N.D., Albers, M.W., Altundag, A., Appenzeller, S., Costanzo, R.M., Croy, I., Davis, G.E., et al. 2022. International consensus statement on allergy and rhinology: Olfaction. Int Forum Allergy Rhinol. 12:327–680.

Pett, M., Lackey, N., and Sullivan, J. 2003. Making Sense of Factor Analysis. 2455 Teller Road, Thousand Oaks California 91320 United States of America: SAGE Publications, Inc.

Putnick, D.L., and Bornstein, M.H. 2016. Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. Developmental Review. 41:71–90.

R Core Team. 2023. A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

Raykov, T. 1997a. Estimation of Composite Reliability for Congeneric Measures. Applied Psychological Measurement. 21:173–184.

Raykov, T. 1997b. Scale Reliability, Cronbach's Coefficient Alpha, and Violations of Essential Tau-Equivalence with Fixed Congeneric Components. Multivariate Behavioral Res. 32:329–353.

Ribeiro, J.C., Simões, J., Silva, F., Silva, E.D., Hummel, C., Hummel, T., and Paiva, A. 2016. Cultural Adaptation of the Portuguese Version of the "Sniffin' Sticks" Smell Test: Reliability, Validity, and Normative Data. PLoS ONE. 11:e0148937.

Rosseel, Y. 2012. **lavaan** : An *R* Package for Structural Equation Modeling. J Stat Soft. 48.

Schiffman, S.S. 1997. Taste and Smell Losses in Normal Aging and Disease. JAMA. 278:1357–1362.

Sorokowski, P., Karwowski, M., Misiak, M., Marczak, M.K., Dziekan, M., Hummel, T., and Sorokowska, A. 2019. Sex Differences in Human Olfaction: A Meta-Analysis. Front Psychol. 10:242.

Steinmetz, H. 2013. Analyzing Observed Composite Differences Across Groups. Methodology. 9:1–12.

Taber, K.S. 2018. The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. Res Sci Educ. 48:1273–1296.

Tan, B.K.J., Han, R., Zhao, J.J., Tan, N.K.W., Quah, E.S.H., Tan, C.J.-W., Chan, Y.H., Teo, N.W.Y., Charn, T.C., See, A., et al. 2022. Prognosis and persistence of smell and taste dysfunction in patients with covid-19: meta-analysis with parametric cure modelling of recovery curves. BMJ. e069503.

Tavakol, M., and Dennick, R. 2011. Making sense of Cronbach's alpha. Int J Medical Education. 2:53–55.

Trentin, S., Oliveira, B.S.F. de, Borges, Y.F.F., and Rieder, C.R. de M. 2022. Evaluation of the complete Sniffin Sticks Test versus its subtests in differentiating Parkinson's disease patients from healthy controls. Arq Neuropsiquiatr. 80:908–913.

Vandenberg, R.J., and Lance, C.E. 2000. A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. Organizational Research Methods. 3:4–70.

Widaman, K.F., and Reise, S.P. 1997. Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In: K.J. Bryant, M. Windle, and S.G. West, eds. The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research. Washington: American Psychological Association. pp. 281–324.

Widaman, K.F., and Revelle, W. 2022. Thinking thrice about sum scores, and then some more about measurement and analysis. Behav Res. 55:788–806.

Wolfensberger, M. 2000. Sniffin'Sticks: a New Olfactory Test Battery. Acta Oto-Laryngologica. 120:303–306.

Wu, S.S., Cabrera, C.I., Quereshy, H.A., Kocharyan, A., D'Anza, B., and Otteson, T. 2023. Olfactory dysfunction incidence and resolution amongst 608 patients with COVID-19 infection.

American Journal of Otolaryngology. 44:103962.