Assessing olfactory abilities with the University of Pennsylvania smell identification test: A rasch scaling approach

Rense Lange[d], Carla L. Donathan[a] and Larry F. Hughes[a,b,c,*]

[a]*Center for Alzheimer Disease and Related Disorders, Southern Illinois University School of Medicine, Springfield, IL, USA*

[b]*Department of Surgery Division of Otolaryngology, Southern Illinois University School of Medicine, Springfield, IL, USA*

[c]*Department of Neurology, Southern Illinois University School of Medicine, Springfield, IL, USA*

[d]*Illinois Department of Education, Southern Illinois University School of Medicine, Springfield, IL, USA*

Abstract: The strategy of delaying or retarding the progression of Alzheimer's disease requires early diagnosis and treatment. Previous research indicates that measurement of changes in olfaction and cognition will play an important role in the early detection of AD and in the monitoring of therapy effectiveness. Using the data of 177 subjects, our objective was to study the measurement properties of the University of Pennsylvania Smell Identification Test (UPSIT) using a Rasch scaling framework. The results indicate that the UPSIT can yield a linear, unbiased, and unidimensional Rasch measure of human smell recognition abilities. As expected, olfactory recognition ability decreased with age, and at the rate of about 0.05 Logits per year. Also, Alzheimer's patients showed a decrease in smell recognition equivalent to that experienced by healthy subjects over the course of 30 years. Hormone replacement therapy was not found to affect healthy women's olfactory recognition ability. Additional diagnostic information can be extracted from the analysis of incorrect responses patterns that is relevant to group membership.

Keywords: Alzheimer's disease, UPSIT, Rasch scaling, item bias, distractor analysis

1. Introduction

Alzheimer Disease (AD) effects more than 4 million people in the USA and the current annual cost of care is estimated to be 115 billion dollars [12,25]. The number of people with AD is projected to more than double by the year 2030. Currently there is no cure for AD and it is unknown when or if one will be available. It appears that one of the best strategies to deal with the growing AD problem is to find therapies that will delay, arrest, or slow the progression of the disease. Unfortunately, AD has an insidious onset and it may take a number of years before any symptoms become evident through casual observation. Furthermore, the progression of the disease is gradual, at least until the late stages of the disease. Different symptoms of AD become apparent in different stages of the disease. As is depicted in figure 1, olfactory and cognitive changes occur early in the course of the disease [see 11 for a discussion of early olfactory deficits] before changes in personality and behavior become evident. Finally, changes in health, that are related to AD, do not occur until relatively late in the course of the disease. If the strategy of delaying or retarding the progression of the disease is to be effective, it is imperative that the diagnosis and treatment be given early in the course

[*] Corresponding author: Larry F. Hughes, Ph.D., Center for Alzheimer Disease and Related Disorders, Southern Illinois University School of Medicine, P.O. Box 19643, Springfield, Illinois, 62794-9643, USA. Tel.: +1 217 545 7186; Fax: +1 217 785 5444; E-mail: lhughes@siumed.edu.

ISSN 1387-2877/02/$8.00 © 2002 – IOS Press. All rights reserved

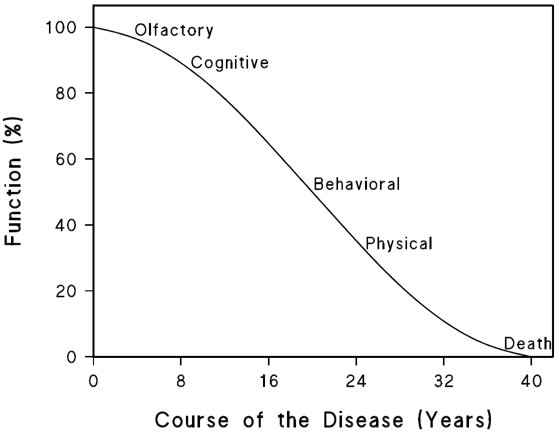

Fig. 1. The theoretical relationship between the course of the disease, the expression of symptoms, and the overall level of functioning. As depicted in the figure, olfactory and cognitive changes occur early in the course of the disease.

of the disease. Thus, the measurement of changes in olfaction and cognition will play an important role in the early detection of AD and in the monitoring of therapy effectiveness.

A number of recent articles related to the use of olfactory tests as diagnostic aids has appeared in the literature [3,5,8,10,11]. In addition the Early Alert TM home-screening test for Alzheimer Disease was released in May of 2001. The vast majority of the olfactory testing is predicated upon the micro encapsulated odors in the University of Pennsylvania Smell Identification Test (UPSIT) [6]. The scoring of these tests typically follows a classical test theory approach [17], which assumes that the number (or percent) of correctly identified stimuli, i.e., the "raw score", can be used as a measure of a person's olfactory identification ability. Unfortunately, raw scores provide only ordinal indices of olfactory identification ability, and such scores are inherently specific to the particular set of olfactory stimuli that were included in the test. Both of these limitations can be solved through the use of Rasch scaling techniques [23,26], which also provides important additional insights into the fundamental measurement properties of the olfactory stimuli. In particular, it can be determined whether the stimuli's measurement properties generalize across different subject groups. Moreover, the lack of generalizability can be exploited to obtain further diagnostic information.

The UPSIT provides a wide variety of olfactory stimuli; hence, it is ideally suited to determine the scalability of human olfactory recognition ability in healthy and AD individuals. If successful, Rasch scaling of the UPSIT will yield cleaner measurements of persons and items, thereby facilitating the implementation of strategies for early diagnosis and management of AD. Although Rasch scaling has become common place in many areas of research [1,7] – including medical applications [22] – this approach has not been used in the present context. For this reason, it is discussed in some detail below.

2. Rasch scaling

A major disadvantage of the traditional raw score approach is that that such scores do not provide linear measures of the underlying trait [23], in this case olfactory identification ability. Using raw scores is like trying to measure the length of several objects with a ruler that has irregular spaced marks (perhaps one end has a number of marks that are closely spaced while the other end has marks that are widely spaced). Different sized objects would be measured with differing degrees of accuracy and comparisons of "length" among objects would not be meaningful. In other words, raw scores are ordinal measures at best, thus distorting group differences and possible treatment effects. Also, the traditional approach essentially treats all items as equivalent, thereby ignoring how the difficulty of the stimuli interacts with persons' olfactory identification ability to produce the outcome of the test [27]. Thus, it is difficult to select those olfactory stimuli that are most appropriate for a particular population of subjects to be tested (we wouldn't think of using a second grade reading test on college freshmen and vice versa). Finally, the standard raw score approach does not recognize that some items may be biased such that subjects with identical olfactory identification ability might receive systematically different scores. This might be the case for instance when women find some olfactory stimuli easier (or harder) to recognize than do men with equal identification abilities as have women. Also, the possibility exists that AD selectively changes people's olfactory identification ability for some stimuli but not others.

These issues are addressed explicitly in Rasch scaling. In particular, the transformed probability P_{ns} that person n correctly identifies UPSIT stimulus s is modeled as a linear function of this persons' position on the latent olfactory identification variable (θ_n) and the difficulty of this stimulus (δ_s):

$$\log\left(\frac{P_{ns}}{1-P_{ns}}\right) = \theta_n - \delta_s. \qquad (1)$$

Note that the person measures θ and the item characteristics δ are in the same metric, i.e., the log-odds ratio (or, logit) in the left hand side of Eq. (1). It can be shown that θ and δ form a linear (i.e., interval level) scale as these two quantities obey the axioms of conjoint additivity [2,21]. Although persons' raw scores (in this case, the number of correctly identified UPSIT stimuli) form a sufficient statistic for respondents' θ values, they are a nonlinear function of respondents θ [27]. Accordingly, only θ provides meaningful estimates of group differences and age trends in olfactory recognition.

Solving for P_{ns} in Eq. (1) yields the expression:

$$P(\theta|\delta_s) = \left(1 + e^{-\theta+\delta_s}\right)^{-1}, \tag{2}$$

which implies that greater δ_s (stimulus difficulty) *decrease* the probability of a correct response while greater θ (identification ability) *increase* this probability. This is illustrated by the two leftmost curves in Fig. 2 which represent the values of Eq. (2) across θ for two hypothetical stimuli A and B, with locations $\delta_A = -1$ and $\delta_B = 1$. It can be seen that $P(\theta|\delta_A) > P(\theta|\delta_B)$ for all θ, i.e., the response curves never overlap. Analogous to ticks-marks on a standard ruler, the item locations δ_s quantify the latent dimension, and it is customary to define them as the point at which $P(\theta|\delta_s) = 0.5$. Also, if the UPSIT is Rasch scaleable, plotting its stimuli along the Logit scale yields an "item map" which provides qualitative information concerning human olfactory recognition.

Figure 2 further illustrates that the Rasch model assumes that the response curves have equal slopes, and that the probability of answering correctly approaches zero (one) for sufficiently low (high) θ. It can be shown [26] that the conditions specified by Eqs (1) and (2) are in fact required to obtain additive (i.e., interval level) person and item measures, and in this case maximum likelihood estimates of θ can be obtained by a nonlinear transformation of persons' raw scores. Using the Bilog software [19] we checked the plausibility of these assumptions by comparing the fit of the Rasch model to that of a competing model that allows for differing asymptotes as well as varying item slopes (see rightmost curve in Fig. 2).

Parameters and fit. We used Linacre's [15,16] Facets software to obtain maximum-likelihood estimates of δ and θ in Eqs(1) and 2. Facets also provides the ingredients for constructing a raw sum to Rasch person measure conversion table, and it quantifies the fit of the items and the persons to the Rasch assumptions based on the discrepancies between the actual and predicted response records [27]. In particular, the *infit* reflects the fit of the items (persons) relative to stimuli (persons) with similar locations, while the *outfit* reflects the items' (persons') fit relative to stimuli (persons) with dissimilar locations. The theoretical value of both statistics is 1. Noisy responses (e.g., due to guessing) tend to increase a stimulus' infit and outfit values, whereas low infit values signify unexpected consistency (e.g., due to the presence of clearly wrong answers). Although fit values from 0.7 to 1.3 are generally deemed acceptable [27], larger ranges of values have been used as well [1].

SE_θ and reliability. Each subject's olfactory ability is estimated 40 times (once for each item) and each item's difficulty is estimated 177 times (once for each subject). The error of measurement SE_θ associated with the person measures decreases to the extent that the θ_n coincide with the item locations δ_s [27]. Except in the presence of severe item bias (see below), the item locations do not vary with the sample being used to estimate them. Accordingly, the SE_θ are essentially sample-free. In practice the variation in item locations tends to be smaller than that of the person locations. As a result, extreme person measures (i.e., high or low) tend to have larger SE_θ than intermediate ones. For instance, although respondents who correctly identify almost all (none) of the stimuli must have a high (low) trait value, we do not know *how* high (or *how* low), and thus SE_θ increases in such cases.

Given that the error of estimate and reliability are just two sides of the same coin [17], the preceding implies that reliability is local as well. In particular, if the SE_θ are treated analogously to the standard error of measurement within classical test theory, the local reliability R_θ is defined as [4]:

$$R_\theta = 1 - \frac{SE_\theta^2}{S_\theta^2}, \tag{3}$$

where S_θ^2 represents the variance of the observed person measures. The presence of S_θ^2 in Eq. (3) implies that in addition to varying with θ, the R_θ are sample specific as well.

Item-level bias. In order for Eqs (1) and (2) to hold in general, the items' properties should be the same across different subgroups of respondents. Thus, given equal smell identification ability, factors such as gender, age, or subject type should not affect the probability with which a particular UPSIT stimulus is correctly identified. (Note that this does not preclude that one subgroup may perform better than another subgroup). Violations of this condition might occur for instance

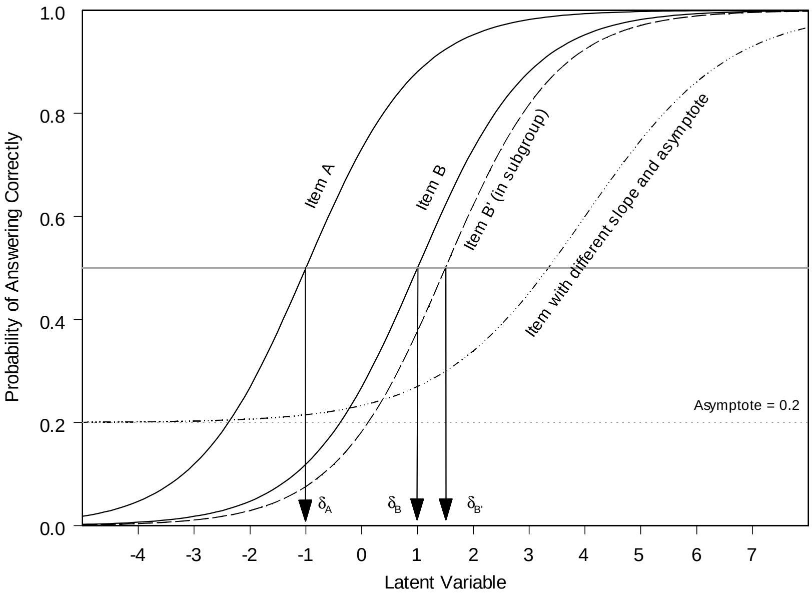

Fig. 2. Hypothetical rasch and non-rasch response curves.

due to group-specific physiological changes that affect subjects' abilities to recognize particular smells. The psychometric literature [17] refers to this type of bias also as Differential Item Functioning, or DIF, and we use these terms interchangeably.

As is illustrated by the second (solid) and third (dotted) curves in Fig. 2, in Rasch scaling the presence of DIF is equivalent to the finding that the location of a stimulus varies across different subgroups (g). Facets provides two types of DIF tests to detect such differences: *First*, the software estimates group specific item locations δ_{sg} so as to obtain bias terms $B_{sg} = \delta_{sg} - \delta_s$ and their standard errors of estimate SE_{sg}. Analogous to the rationale behind the t-test, an approximate statistical test for the case of two subgroups g_1 and g_2 is obtained by assuming that the B_{sg} are normally distributed:

$$z_{s(g1)(g2)} = \frac{B_{s(g1)} - B_{s(g2)}}{\sqrt{SE^2_{s(g1)} + SE^2_{s(g2)}}}. \qquad (4)$$

Secondly, summing all values $z^2_{Bsg} = (B_{sg}/SE_{sg})^2$ yields an omnibus χ^2 test (with df = number of items × number of groups) that considers all group specific item locations simultaneously.

Test-level bias. Biases in individual items may combine to introduce systematic biases at the test level as well. If so, the raw sum to Rasch person measure transformation will vary across subgroups of respondents. Conversely, when these transformations are similar across subgroups of respondents one may assume that the item level biases (if any) tended to cancel.

Dimensionality. Basic Rasch scaling assumes that the stimuli and the respondents vary along a single dimension. Although the finding of acceptable item fit supports this assumption [9], a more stringent test is obtained by comparing the relative fit of the intended unidimensional model to that of a suitable competing two-dimensional model. Such tests can be obtained via the ConQuest Rasch software [28].

3. Method

3.1. Subjects

Most of the healthy subjects were women that had participated in a study of the effects of hormone replacement therapy (HRT) on olfaction. The study included 62 women recruited from obstetrics-gynecology clinics at SIU ranging in age from 41.56–76.01 who had self-selected either HRT (varying types) or no HRT. They scored in the cognitively normal range on the Mini Mental Status Exam (MMSE) at the time of pre-study evaluation, and they provided signed, dated, and witnessed informed consent approved by the institutional review board. All subjects considered for inclusion in this study received a medical screening. Screening consisted of a complete medical history, physical examination (including blood pressure, height, and weight), a gynecologic examination and cervical cytologic smear (if not performed within the past year). Serum follicle stimulating hormone (FSH) levels were determined at the time of initial screening and physical. Both opposed and unopposed HRT treatments were included in this group.

The remaining subjects were patients that had been referred to the Southern Illinois University School of Medicine Center for Alzheimer Disease and Related Disorders (CADRD) to be screened for AD. Olfactory testing is routinely done as a part of the initial assessment. The data represent the intake of the dementia clinic at CADRD for the last nine months.

Taken together, this study involved 177 subjects [31 men (M) and 146 women (F)] ranging in age from 39.57 to 89.69 years. Of these, 48 (31 F, 17 M) were diagnosed with Alzheimer's disease, 15 (15 F, 0 M) with Parkinson's disease, and a mixed group of 41 (28 F, 13 M) subjects had other diagnoses, including mild cognitive impairment, progressive aphasia, migraine, and uncertain diagnoses. Finally, 73 (72 F, 1 M) individuals were included as a control group. The control group included the 62 women from the HRT study.

3.2. *Materials and procedure*

The UPSIT was administered in accordance with the test instructions. However, the experimenter read each stimulus' four response choices to the subjects, in the order in which they occurred, both before and after they scratched and sniffed the odorant.

4. Results

4.1. *UPSIT scaling properties*

All items and persons were scaled simultaneously using the Facets software, yielding stimulus locations ($M = -1.32$, $SD = 0.79$) and person measures ($M = 0.07$, $SD = 1.55$) expressed in a common Logit metric. Figure 3 summarizes the most important properties of these two quantities. The bottom section of this figure shows the approximate distribution of the stimulus locations (their exact values and standard error of estimate are shown in Columns 1 and 2 of Table 1), together with the 40 person measurement values that are defined by the UPSIT items (triangles on X-axis). The person measures' smoothed densities in the four subject groups are shown in the middle of Fig. 3.

It can be seen that the distributions in the Alzheimer and Parkinson conditions center on the item locations. Accordingly, as is indicated by the top curve, the local reliability of the person measures is optimal for these two groups. By contrast, the UPSIT items are too easy for many Healthy subjects as their person measures (and those of several Mixed subjects) exceed the locations of the smell stimuli. Note that this greatly decreases the precision with which Healthy subjects can be measured (i.e., the triangles on the X-axis are spaced further apart), as well as the local reliability of their measures. Thus, the quality of the measures obtained for the Healthy subjects is inferior to that of the AD and Parkinson subjects.

The reliability of the UPSIT as determined within the framework of classical test theory appears excellent (*KR-20* $=$ 0.93). Also, only 19 of the 177 respondents (i.e., about 11%) show outfit values that deviate significantly from the optimal value 1 at $p < 0.10$. Since the number of misfitting subjects barely exceeds what is to be expected by chance alone (i.e., 17.7), it follows that the subjects behaved in accordance with the Rasch assumptions. Nevertheless, the behavior of several stimuli is problematic. For instance, Table 1 (Columns 3 and 4) lists the stimuli's infit and outfit statistics. As is shown by the boldface entries, the infit of three stimuli (i.e., 12, 14, and 27) exceeds 1.3. Also, the outfit values of smells 8, 9, 13, 24, 29, 34, 35, and 40 fall below 0.7, indicating that the responses to these stimuli are too predictable relative to the answers for other stimuli. By contrast, the responses to stimuli 7, 12, 14, 27, and 36 are too noisy (erratic) as their outfit exceeds 1.3.

Item-misfit can have a variety of causes, including the prevalence of non-Rasch response curves (item characteristic curves that do not meet the requirements for implementing the Rasch model, see below), response biases, and multidimensionality. Our attempts to narrow down the possibilities proved to be instructive and we report the findings in some detail.

The three-parameter logistic. Perhaps indicative of guessing by low performing subjects and/or carelessness of high performing subjects, the UPSIT items' outfit values increase significantly with their Rasch locations (*Kendall's tau* $=$ 0.53, $p <$ 0.001). Also, the finding of large infit values suggests that the slopes of the stimuli's response curves vary across items. Pervasive guessing by respondents, as well as differently sloped response curves, contradict the assumptions of the Rasch model. Hence, using the Bilog software [19], we tested the fit of the Rasch formulation against that of the most general Item-Characteristic formulation, the

Table 1
Summary of rasch analyses

Stimulus No.	Stimulus (correct answer, followed by distractors)[a]	UPSIT-40					UPSIT-29				
		δ (1)	SE_δ (2)	Infit (3)	Outfir (4)	AD vs. control z (5)	δ (6)	SE_δ (7)	Infit (8)	Outfir (9)	AD vs. control z (10)
1	PIZZA (gasoline, peanuts, lilac)	−0.89	0.19	0.8	0.7	0.30	−0.46	0.21	0.9	0.9	−0.88
2	BUBBLE GUM (dill pickle, wintergreen, watermelon)	−0.74	0.19	1.0	1.0	−0.99	−0.29	0.21	1.2	1.3	−2.39
3	MENTHOL (tomato, licorice, strawberry)	−1.88	0.21	1.0	1.0	−0.89	−1.59	0.22	1.1	1.2	−2.04
4	CHERRY (whiskey, honey, lime)	−1.93	0.22	0.9	0.7	1.49	−1.64	0.23	1.0	1.3	0.73
5	MOTOR OIL (grass, pizza, pineapple)	−2.28	0.23	0.9	0.7	1.14	−2.02	0.24	0.9	0.8	0.38
6	MINT (skunk, fruit punch, cola)	−1.22	0.20	1.0	0.9	−0.52	−0.85	0.21	1.1	1.2	−1.78
7	BANANA (garlic, cherry, motor oil)	−1.46	0.20	1.1	**1.5**	−1.98					
8	CLOVE (licorice, chili, banana)	−2.07	0.22	0.9	**0.6**	1.14	−1.80	0.23	0.9	0.8	0.38
9	LEATHER (clove, lilac, apple)	−2.44	0.24	0.8	**0.6**	1.14	−2.20	0.25	0.9	0.8	0.57
10	COCONUT (skunk, cedar, honey)	−0.96	0.19	0.9	1.1	0.83	−0.54	0.21	1.1	1.2	−0.35
11	ONION (chocolate, banana, fruit punch)	−1.62	0.21	0.9	0.8	2.05	−1.30	0.22	1.0	1.2	1.29
12	FRUIT PUNCH (soap, menthol, pumpkin pie)	0.51	0.19	**1.4**	**1.8**	−**3.75****					
13	LICORICE (pineapple, cheddar cheese, cherry)	−1.42	0.20	0.8	0.6	1.01	−1.07	0.21	0.8	0.7	0.02
14	CHEDDAR CHEESE (paint thinner, cherry, coconut)	0.01	0.18	**1.4**	**1.5**	−1.16					
15	CINNAMON (cola, pine, coconut)	−0.89	0.19	1.1	1.3	−**2.73***					
16	GASOLINE (rose, lemon, peach)	−2.97	0.27	1.0	0.8	0.32	−2.75	0.28	1.1	1.1	−0.41
17	STRAWBERRY (dill pickle, chocolate, cedar)	−1.34	0.20	0.9	0.8	0.96	−0.98	0.21	1.0	1.2	−0.13
18	CEDAR (gasoline, lemon, root beer)	−1.38	0.20	1.1	1.2	0.57					
19	CHOCOLATE (lemon, root beer, black pepper)	−2.03	0.22	0.9	0.7	0.89	−1.74	0.23	1.0	0.9	0.14
20	GINGERBREAD (menthol, apple, cheddar cheese)	−0.19	0.18	1.2	1.3	−**3.13***					
21	LILAC (chili, coconut, whiskey)	−2.07	0.22	0.9	0.7	−0.37	−1.80	0.23	1.0	0.9	−1.36
22	TURPENTINE (soap, skunk, chili)	−0.40	0.19	1.1	1.0	−1.66					
23	PEACH (chocolate, leather, pizza)	−1.14	0.19	1.1	1.1	0.89					
24	ROOT BEER (watermelon, banana, smoke)	−1.42	0.20	0.7	**0.5**	1.51	−1.07	0.21	0.8	0.6	0.62
25	DILL PICKLE (pineapple, root beer, black pepper)	−1.22	0.20	0.8	0.9	0.30	−0.85	0.21	0.9	0.9	−0.88
26	PINEAPPLE (smoke, whiskey, onion)	−1.07	0.19	1.0	0.8	0.79	−0.67	0.21	1.1	1.3	−0.37
27	LIME (musk, garlic, turpentine)	0.21	0.18	**1.4**	**1.6**	−**4.13****					
28	ORANGE (cheddar cheese, bubble gum, turpentine)	−1.03	0.19	1.0	0.9	0.94	−0.63	0.21	1.1	1.1	−0.13
29	WINTERGREEN (lime, pumpkin pie, leather)	−1.75	0.21	0.9	**0.6**	1.93	−1.44	0.22	1.0	0.7	1.17
30	WATERMELON (chili, menthol, orange)	−2.28	0.23	0.9	0.7	−0.06	−2.02	0.24	0.9	0.9	−0.93
31	PAINT THINNER (watermelon, peanut, rose)	−2.28	0.23	1.0	0.8	0.62	−2.02	0.24	1.1	0.9	−0.13
32	GRASS (mint, gingerbread, strawberry)	−0.81	0.19	1.1	1.2	−0.37					
33	SMOKE (dill pickle, grass, peach)	−0.64	0.19	0.9	0.8	**3.02***	−0.16	0.21	1.1	1.1	2.03
34	PINE (smoke, lilac, orange)	−1.38	0.20	0.8	**0.6**	1.95	−1.03	0.21	0.9	0.7	1.17
35	GRAPE (pizza, turpentine, clove)	−2.23	0.23	0.9	**0.5**	1.43	−1.96	0.24	1.0	0.6	0.86
36	LEMON (motor oil, pumpkin pie, rose)	0.11	0.18	1.3	**1.5**	−**3.07***					
37	SOAP (black pepper, licorice, peanut)	−1.54	0.20	0.9	0.8	1.71	−1.21	0.22	1.0	1.2	0.95
38	NATURAL GAS (orange, musk, cola)	−1.54	0.20	1.1	0.9	−0.55	−1.21	0.22	1.2	1.3	−1.66
39	ROSE (lime, mint, bubble gum)	−0.85	0.19	0.8	0.7	**2.68***	−0.42	0.21	0.9	0.8	1.67
40	PEANUT (lemon, apple, root beer)	−2.28	0.23	0.8	**0.5**	1.38	−2.02	0.24	0.8	0.7	1.04

[a] The text does *not* reflect the order of presentation of the stimuli by the experimenter.
$^{*}p < 0.01$; $^{**}p < 0.001$.

three-parameter logistic (3PL). As was already illustrated in Fig. 1, the 3PL generalizes the Rasch model by incorporating an item specific lower asymptote (or, guessing parameter) and a slope parameter.

Not surprisingly, the 3PL showed a better fit to the data (*-2Log Likelihood* $= 3265.29$, $df = 120$) than did the Rasch model (*-2Log Likelihood* $= 3390.99$, $df = 40$). However, this difference in fit (i.e., 125.70) is far less than twice the associated degrees of freedom ($120 - 40 = 80$). Hence, there is no reason to reject the Rasch formulation [19].

Item-level biases. Because item misfit can be the result of group specific response biases, we performed extensive tests for differential item functioning. Facets' omnibus test detected no overall gender ($\chi^2_{80} = 66.9,\ p > 0.50$) or age ($\chi^2_{120} = 119.7,\ p > 0.48$) related DIF, indicating that the stimuli's locations differ little across the levels of these two factors. (Note: Three age groups of approximately equal size were used). Accordingly, men and women, and younger, medium, older respondents *with equal* (*latent*) *abilities to recognize smells* have similar probabilities of

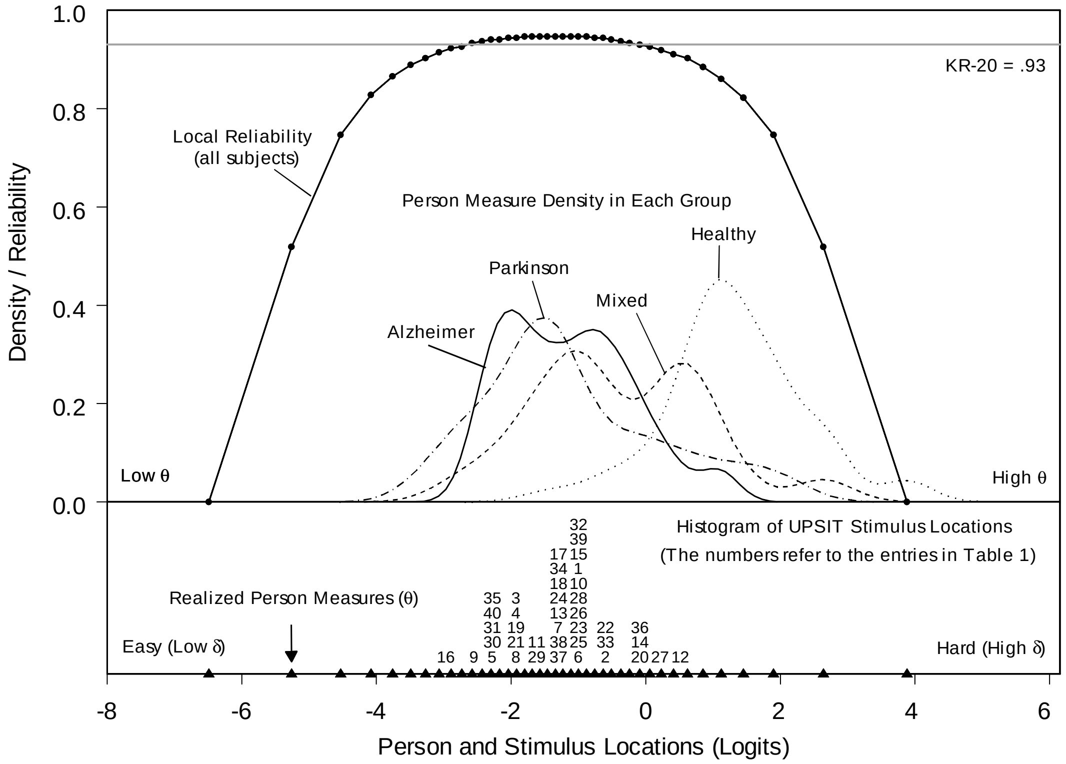

Fig. 3. Distributions of item and person locations.

correctly identifying the UPSIT stimuli.

However, the locations of the olfactory stimuli differ significantly across the four different subject groups ($\chi^2_{160} = 204.4$, $p < 0.001$). Closer inspection of the bias terms B_{sg} (see Eq. (4)) revealed that the observed group effect revolves around differences between the AD and the Healthy groups, and follow-up tests identified seven stimuli as significantly biased ($p < 0.01$, 2-sided, see Column 5 of Table 1). Rather surprisingly, the bias is such that AD subjects performed better than Healthy subjects with equal smell recognition abilities on five of these seven items (i.e., 12 $=$ Fruit Punch, 15 $=$ Cinnamon, 20 $=$ Ginger, 27 $=$ Lime, and 36 $=$ Lemon), and worse on just two (33 $=$ Smoke and 39 $=$ Rose). With the exception of the Smoke stimulus (see Ancillary Analyses), extensive analyses of these stimuli and their distractors provided no definite clues concerning the reasons behind these biases. We further note that all UPSIT kits had the same lot number, thus group specific variations in the manufacturing of the smell stimuli can be excluded as a confounding variable.

Test-level bias. It should be stressed that the item-level biases had little effect at the test level. For instance, Fig. 4 shows the estimated person measures (Y-axis) in these two AD and Healthy groups as a function of the number of correctly identified smells (X-axis). In addition, the vertical dotted lines indicate the standard errors ($\mathrm{SE}_{\mathrm{dif}}$) associated with the difference between the two estimates relative to their averaged value. It can be seen that the two curves are quite close and that their difference never exceeds $\mathrm{SE}_{\mathrm{dif}}$. Hence, the biases in the individual UPSIT stimuli tended to cancel at the test level. As a result the net distortion of the person measures is negligible. As an aside we note that Fig. 4 also clearly reveals the nonlinear relation between raw scores and actual person measures as determined by maximum likelihood methods.

A purified UPSIT. It might seem that the UPSIT can be improved by excluding some ill-fitting or biased stimuli. However, since the ill-fitting stimuli are also among the most difficult ones, their removal might further degrade the measurement of high performing subjects. This is confirmed by additional Facets runs in which all misfitting and biased items were removed by an iterative Top-Down purification procedure [13,14]. As is shown in Columns 6 through 10 of Table 1, 11 items were removed, leaving 29 items. We note that removing poorly performing items often induces misfit or bias in the remaining items. Hence, some of the stimuli

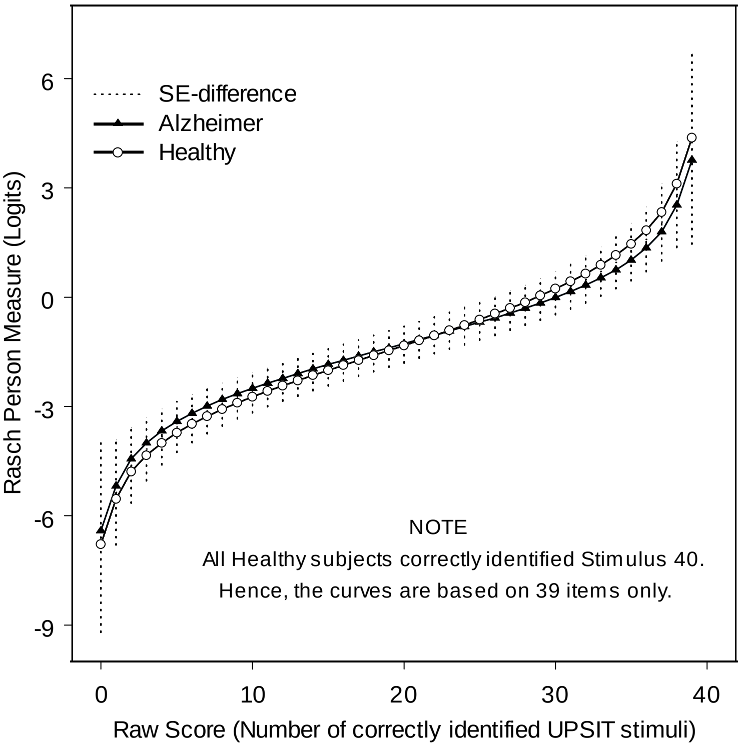

Fig. 4. Raw score to rasch person measure translations in the Alzheimer and healthy subject groups.

that initially appeared acceptable (see Columns 3 and 4 of Table 1) were removed as well.

Psychometric theory [9,24] and computer simulations [13] alike indicate that item misfit and DIF may be the result of multidimensionality. To test whether the sub-optimal properties of the UPSIT are due to multidimensionality, the 29 items of the purified UPSIT and the 11 sub-optimal items were treated as two separate Rasch factors. Using the ConQuest software [28], the fit of this two-factor formulation was compared to that of a single factor model that comprises all 40 UPSIT items. Although the two-factor model showed a statistically superior fit ($\chi^2_2 = 108.09,\ p < 0.001$), the direct (i.e., attenuation-corrected) correlation between the two factors is extremely high ($r = 0.96$). Since the magnitude of this correlation makes any distinction between the two factors rather meaningless, a one-factor formulation remains to be preferred.

Evaluation. All results that are reported in following section are based exclusively on the 40-item Rasch version of the UPSIT. This choice reflects our opinion that the restriction of range introduced by the 29-item version is deemed as more serious than the imperfections of the longer version. Table 2 shows the raw-score to Rasch translation of the 40-item UPSIT together with the local standard errors of measurement (in Logits).

A qualitative interpretation of the Rasch measure is provided by the item map in Fig. 5, which plots the locations of the UPSIT stimuli along the vertical center axis. We note that the stimulus locations' error of measurement is about 0.2 Logits (see also Table 1), i.e., pairwise differences less than 0.6 Logits are probably meaningless (this value corresponds to twice the (rounded) standard error of difference between locations). Not surprisingly, gasoline and leather are the most recognizable; followed by other distinctive odors such as watermelon, paint thinner, grape, clove and motor oil. By contrast, fruit punch, cheddar cheese, lime, lemon, ginger and turpentine require the greatest smell recognition ability. The remaining stimuli form a densely spaced middle group with similar difficulty levels. The leftmost "ruler" describes how the probability by which an item is endorsed (right side) depends on the difference $\theta - \delta$ (in Logits) between the locations of items and persons (left side). For instance, a subject with $\theta = -0.96$ has a modeled probability of 50% of correctly recognizing the UPSIT Coconut stimulus, as $\delta_{\text{Coconut}} = -0.96$. To increase this probability to 67%, 0.71 more Logits of smell recognition are required. Since all items have identical Rasch curves (except for offset, see Eq. (2)), similar predictions for

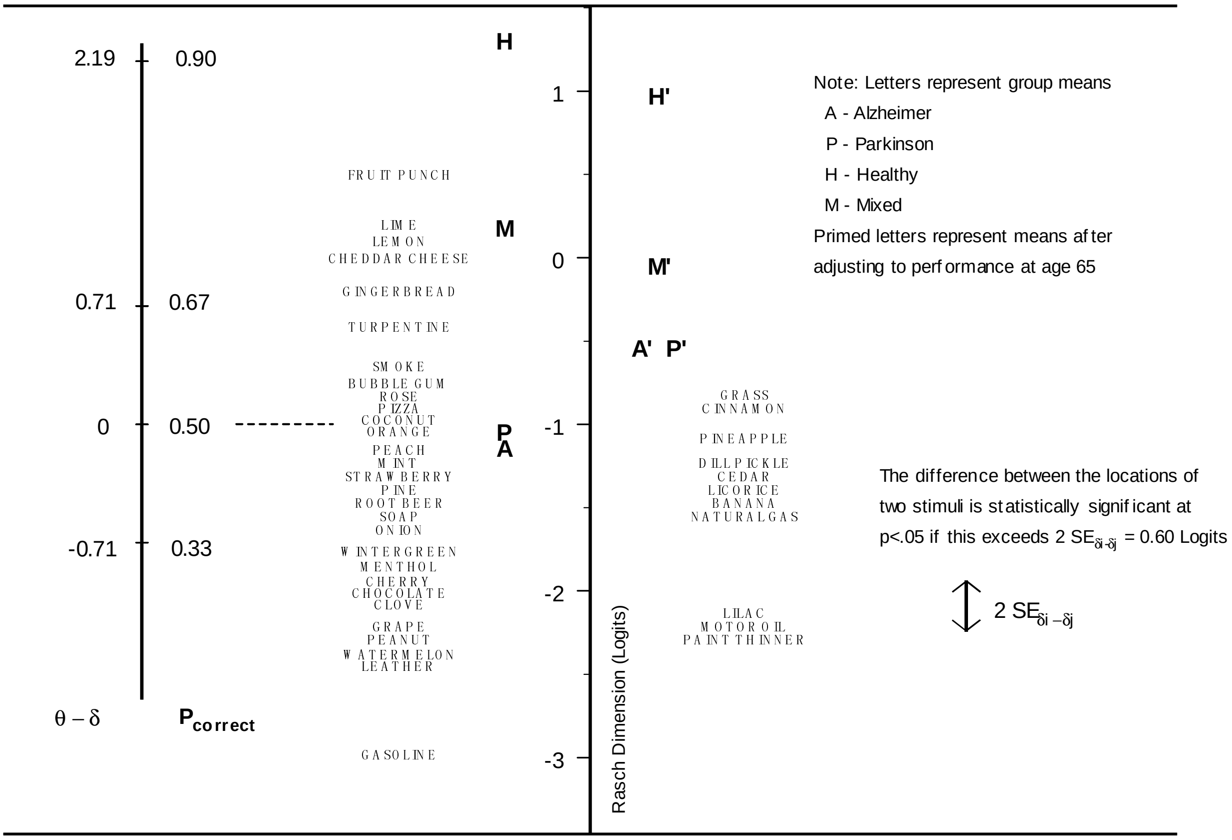

Fig. 5. Item map of UPSIT olfactory stimuli.

other items can be obtained by centering the ruler at these items' positions.

No simple interpretation can be given to the smell recognition dimension as a whole. For instance, one might hypothesize that many natural odors such fruit punch, lime, lemon, grass, rose, smoke and orange, are harder to recognize than the odors of man-made products such as gasoline, paint thinner, and motor oil. However, there are several exceptions that contradict this generalization (e.g., grape and lilac vs. turpentine). We further note that a stimulus' ease of recognition depends partly on the attractiveness of the incorrect choices (distractors) that are presented to the respondent (see Table 1). It is possible therefore that a different hierarchy would emerge if different distractors were used.

4.2. Group differences

It is clear from Fig. 3 that Healthy subjects perform considerably better on the UPSIT than do the other three subject types. However, the effects in this figure are confounded, as subject age is highly predictive of UPSIT performance ($r = -0.70,\ p < 0.001$), while the median age in years of the Healthy subjects (58.0) is much lower than that of the AD (76.5), Parkinson (78.0) and Mixed (76.0) subjects (Kruskal-Wallis, $\chi^2_3 = 77.33,\ p < 0.001$). For these reasons, age was used as a covariate in a 4 (Group: AD, Parkinson, Healthy, Mixed) Analysis of Covariance (ANCOVA) over the UPSIT person measures in Logits. (Note: Because the Parkinson and Healthy groups contained 0 and 1 men, respectively, gender was not used as an independent variable).

The data did not contradict the ANCOVA assumptions ($F(1,89) = 0.87$, n.s.), and the effects of Age ($F(1,172) = 41.36,\ p < 0.001$, *partial* $\eta^2 = 0.19$) and subject group ($F(3,172) = 15.71,\ p < 0.001$, $MS_e = 0.99$, *partial* $\eta^2 = 0.22$) were highly significant. Note that the η^2 values reflect that age and subject group contribute nearly equally to smell recognition. As is also shown in Fig. 5, when evaluated for age 65, the marginals in the AD, Parkinson, Healthy, and Mixed groups are -0.52, -0.52, 0.99, and -0.03 Log-

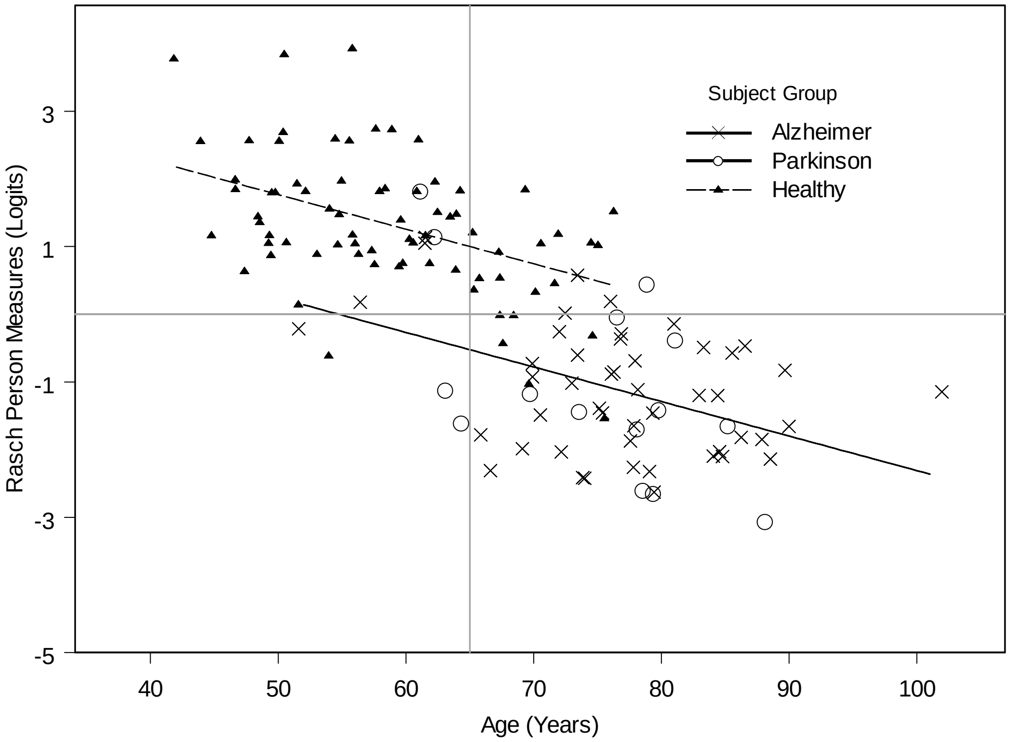

Fig. 6. Person measure in the Alzheimer, Parkinson, and healthy groups by age.

its, respectively. Pairwise comparisons using the Least Significant Difference Method indicate that Healthy subjects performed significantly better than the other three subject groups ($p < 0.001$). All other group comparisons failed to reach statistical significance at $p < 0.01$. Figure 6 shows the person measures of the AD, Parkinson, and Healthy subjects by age. Note that the measures of the first two subject groups are completely intermixed, a finding that is consistent with the literature and a recent meta-analysis of olfaction in AD and Parkinson [18]. Thus, the similar means for the Alzheimer and Parkinson groups are not due to outliers that somehow cancelled in the aggregate.

Ancillary analyses. To determine whether the preceding results are due the large proportion of women in the Healthy group, we performed a 2 (Gender) × 2 (Group: AD, Mixed) ANCOVA over the Rasch measures using age a covariate. Women's Rasch measures exceeded those of the men by 0.45 Logits, but this difference is not statistically significant ($F(1,84) = 2.14,\ p > 0.10$). Also, gender did not interact with subject group ($F(1,84) = 2.83,\ p > 0.09$). Next, the ANCOVA over all four groups was repeated after subtracting the gender effect from the person measures of all women in the Healthy group (i.e., we act as if this groups consist of men only). Despite this conservative correction the group main effect persists ($F(3,172) = 13.12,\ p < 0.001$). Also, pairwise comparisons indicate that the marginal of the Healthy group remains significantly higher than that in the AD, Parkinson, and Mixed groups (all $p < 0.001$). In other words, the superior performance obtained in the Healthy group cannot solely be explained as a gender confound.

Estrogen. A subset of 54 of the 146 women was receiving estrogen treatment at the time of this research. Again using age as a covariate, we performed a 4 (Group: AD, Parkinson, Healthy, Mixed) by 2 (Estrogen: Yes vs. No) ANCOVA over all women's person measures. Receiving estrogen treatment did not affect women's overall smell recognition ($F(1,137) = 0.00$, n.s.). Also, the estrogen factor did not interact with group factor ($F(3,137) = 0.54$, n.s.). A detailed report of the HRT study in which these women participated has been accepted for publication in the Journal of the International Menopause Society (Climateric) 2002.

Quantitative effects. We note that the common slope of the regression line in the four subject groups is -0.053 ($SE = 0.008$) – i.e., at least in the age range studied here, the latent smell recognition variable de-

Table 2
Raw score to rasch person measure transformation for the forty-item UPSIT

Raw	Logits	
sum	Person measure	SE
0[a]	−6.49	1.84
1	−5.26	1.02
2	−4.53	0.74
3	−4.08	0.61
4	−3.76	0.54
5	−3.49	0.49
6	−3.27	0.46
7	−3.07	0.43
8	−2.89	0.41
9	−2.73	0.40
10	−2.58	0.38
11	−2.43	0.37
12	−2.30	0.36
13	−2.17	0.36
14	−2.04	0.35
15	−1.92	0.35
16	−1.80	0.34
17	−1.69	0.34
18	−1.57	0.34
19	−1.46	0.34
20	−1.34	0.34
21	−1.23	0.34
22	−1.11	0.34
23	−1.00	0.34
24	−0.88	0.34
25	−0.76	0.35
26	−0.63	0.35
27	−0.51	0.36
28	−0.37	0.37
29	−0.24	0.38
30	−0.09	0.39
31	0.06	0.40
32	0.23	0.42
33	0.41	0.44
34	0.62	0.46
35	0.85	0.50
36	1.12	0.55
37	1.45	0.62
38	1.90	0.74
39	2.64	1.02
40[a]	3.88	1.84

[a]The entries in this row are extrapolated values, as person measures cannot be estimated when all stimuli are identified correctly or incorrectly. Additionally, very low scores likely reflect malingering, however, to simplify the current discussion, this is not addressed.

creases by about one-half Logit every ten years. Thus, the lower performance exhibited by the AD subjects relative to the Healthy group (i.e., about 1.5 Logits) approximately corresponds to the decrease in smell recognition that Healthy subjects experience over the course of 30 years. Also, he difference between genders was 0.45 Logits. In other words, if in Doty et al.'s [6] original work on the UPSIT one were to shift the age × gender curves out by ten years, the curves for male and females would completely overlap in the age range studied.

4.3. Analyses of distractor choices

Discriminant analyses. We noted earlier that deviations from the Rasch assumptions may provide important additional diagnostic insights. In this context, the finding of powerful DIF effects in the UPSIT stimuli (see Item-Level Biases) suggests that equal performing Alzheimer and Healthy respondents arrive at systematically different interpretations of the same chemical stimuli. Such differences likely reflect that Healthy and AD subjects have differential preferences for the UPSIT stimuli's *incorrect* choices (i.e., the distractor smells). Meaningful study of differential distractor preferences requires that the effects of confounding factors be eliminated as much as possible.

To control for the powerful performance difference between the AD and Healthy group, only subjects with raw scores inside the range 18 to 27 were used (no Healthy subject answered fewer than 18 item correctly and no AD subject answered more than 27 correctly). However, the Healthy subjects are also considerably younger (Range: 45 to 76 years) than the AD subjects (Range: 52 to 101 years). Unfortunately, as only 69 subjects remain after equating for performance, equalizing the age ranges as well would leave too few subjects for analysis. Hence, in the following, subject group (AD vs. Healthy) is necessarily confounded with subject age.

To avoid spurious results, 62 of the 120 distractors were eliminated as they occurred fewer than 5 times over all subjects. The remaining 58 distractors were transformed into dummy coded variables (0 = absent, 1 = present) and subjected to a discriminant analysis using forward stepwise elimination ($p < 0.05$). A single discriminant function resulted ($\chi^2_6 = 71.78$, $p < 0.001$, Canonical Correlation = 0.82) that is defined by the six distractors listed in Table 3. Of the 69 subjects 62 (89.9%) are classified correctly. This number drops very little (to 59, or 85.5%) when jackknife cross-validation is applied (7 AD and 3 Healthy subjects are misclassified). The negative sign of the standardized canonical discriminant function coefficient for the Turpentine distractor (i.e., -0.31) reflects that Healthy (or younger) subjects are more likely to select this distractor when presented with the UPSIT Lime smell. However, the remaining coefficients are all positive. In other words, identifying Grass as Strawberry, Strawberry as Chocolate, Peach as Leather, Grape

Table 3
Discriminant function derived from stimuli's distractors only

UPSIT stimulus #	Smell	Distractor	Frequency of occurrence	Standardized canonical coefficient	Asymmetric lambda
27	Lime	Turpentine	55	−0.31	0.07
32	Grass	Strawberry	18	0.44	0.17*
17	Strawberry	Chocolate	22	0.46	0.10
23	Peach	Leather	33	0.56	0.20*
35	Grape	Clove	13	0.76	0.10
34	Smoke	Peach	16	0.99	0.53**

$^*p < 0.05$; $^{**}p < 0.001$.

as Clove, and Smoke as Peach all are characteristic of the AD (or older) subject group.

Although the six distractors discriminate rather well between Healthy and AD subjects when taken together, most distractors perform rather poorly in isolation. As is indicated by the asymmetric Lambda values in the last column of Table 3, only three combinations (Grass-Strawberry, Strawberry-Chocolate, and Smoke-Peach) have statistically significant predictive values ($p < 0.05$) when subject group is taken as the dependent variable. However, of these three pairs only the Smoke-Peach combination has a non-trivial Lambda value (0.53). In other words, knowing that a subject mistook the UPSIT Smoke stimulus for Peach increases our ability to predict his or her group membership (AD vs. Healthy) by about 53% relative to that afforded by the marginal group frequencies only. For this reason, we studied the Smoke stimulus and its distractors in greater detail.

Smoke-peach. Figure 7 plots the observed probabilities of the Smoke distractors (Dill Pickle and Grass differed little and they were combined for greater clarity) as well as the correct choice for Healthy (top) and AD subjects (bottom). Additionally, a distinction is made between High performing subjects (raw score range: 18–27) and Low performing ones (range: 28–36).

Consistent with the item-level bias reported earlier, Healthy subjects are much more likely to identify this stimulus correctly than are the AD subjects ($\chi^2_1 = 28.67,\ p < 0.001$). Interestingly, none of the 39 Healthy (younger) subjects identified the smoke smell as peach, but the majority of AD (older) subjects (16 out of 30) preferred this incorrect answer ($\chi^2_1 = 27.08,\ p < 0.001$). The choices in the two groups appear to be intentional, as the proportion of peach choices for AD subjects (0.53) exceeds the guessing level, i.e., 0.25 (Binomial Distribution, $p < 0.001$). The corresponding proportion for Healthy subjects (0.0) falls below 0.25 ($p < 0.001$). Finally, the effect is not related to subjects' overall performance on the UPSIT, as Fig. 7 shows that the proportion of Peach choices in both groups does not vary across the two performance levels.

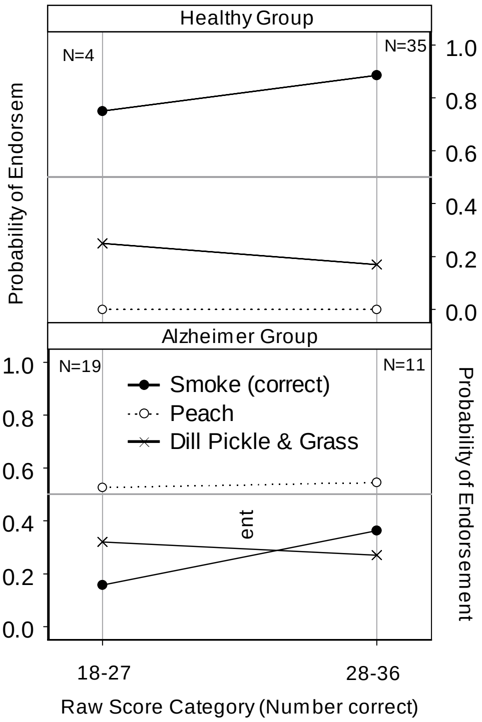

Fig. 7. Probability of endorsement of the smoke stimulus and its distractors as a function of the raw scores in Alzheimer and healthy groups.

Order effects. Inspection of the data suggested that when AD and healthy subjects responded incorrectly, they exhibited different preferences depending on the distractors' positions in the UPSIT test booklet. However, the analysis of simple counts would give a distorted picture since respondents' preference for particular incorrect choices likely varies with their θ. To account for this variable we computed the Rasch locations of the 120 incorrect choices in the UPSIT relative to those of the 40 correct choices. This was done by treating subjects' response records as if these consisted of 160 observations while forcing the correct choices

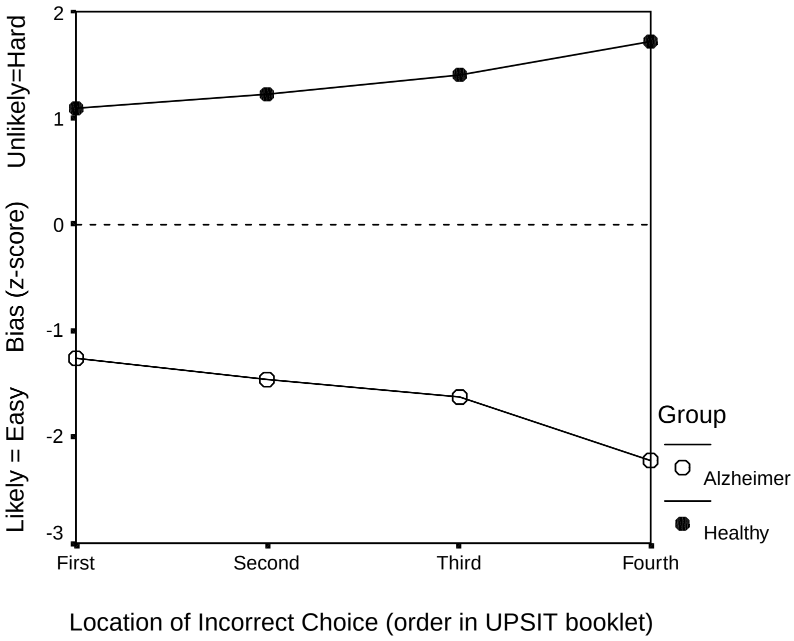

Fig. 8. Average normalized bias terms of distractors as a function of their order of presentation in UPSIT booklet for Alzheimer and healthy respondents.

to assume the locations shown in Table 1. Next, each choice's ordinal position in the UPSIT booklet (i.e., First, Second, Third, or Fourth) was introduced as a factor in Eq. (1) (recall that this was the order in which the experimenter read the choices to the subjects). To determine whether the answer choices were affected by their positions, standardized bias terms were computed across the four positions for AD and healthy subjects.

An analysis of variance by subject group and distractor position over the standardized bias terms showed no main effect of position $(F(3,230) = 0.73)$. Not surprisingly, the main effect of subject group $(F(1,230) = 1238.20,\ p < 0.001)$ was highly significant since the lower performing AD subjects found incorrect choices "easier" ($M = -1.64$ Logits) than did the higher performing healthy subjects ($M = 1.36$ Logits). Most importantly, a subject group by distractor position interaction effect was found $(F(3,230) = 16.34,\ p < 0.001)$. As is indicated by the lower line in Fig. 8, this interaction reflects that AD subjects find distractors later in the sequence increasingly "easy" (i.e., they favored later distractors over earlier ones), while healthy subjects (top line) found the later distractors increasingly "hard" (i.e., they favored earlier distractors over later ones). Post-hoc comparisons in each group across the distractors' ordinal position indicated that the interaction is mainly due to differential preferences for the fourth distractor ($p < 0.05$).

Again, it must be kept in mind that the distinction between AD and healthy subjects is confounded by age. We hypothesize however that the 'primacy effect observed for healthy subjects' is due to a lack of motivation to process all possible choices, thus leading to premature answer selection. By contrast, the "recency effect" exhibited by the AD subjects may be due to a loss of short-term memory such that earlier distractors are already forgotten when these subjects are asked to state their response to the experimenter.

5. Summary and discussion

The "baby boomers" are going to be the healthiest group of people ever to hit their 70s and 80s. Ironically, this will mean more and more people will live long enough to get Alzheimer Disease. There are currently 34.8 million Americans 65 years and older, 4.3 million of whom are 85 or older. Dementia strikes 3% of people aged 65 to 74, 19% of those 75 to 84, and up to 47% of people 85 and older. The number of people with Alzheimer's is expected to more than double over the next 30 to 40 years. Recent longitudinal studies on olfaction and cognitive decline [5,8] have indicated that olfactory dysfunction is an early predictor of those that will ultimately be diagnosed as having AD. It is imperative that our instruments for measuring olfaction and cognition be thoroughly evaluated and refined if they are to be effective in detecting early and subtle changes.

Our findings indicate that the UPSIT can play an important role in this respect, as this instrument provided linear (i.e., interval level) and unbiased measures of human smell recognition abilities. Specifically, subjects' responses to the UPSIT stimuli agreed reasonably well with the assumptions of the Rasch model, and it follows therefore that the UPSIT defines a single quantitative dimension on which respondents can be located with known reliability. Consistent with the research referred to above, olfactory recognition ability decreased with age, and Alzheimer's patients showed significantly lower recognition ability. Given the absence of significant age and gender biases in the overall UPSIT measures, these two effects can now precisely be quantified. First, at least for the age range studied here, the age decrease is about one-twentieth of a Logit per year (i.e., the log-odds of correctly identifying a particular UPSIT stimulus decreases by 0.05 annually). Second, after controlling for age, the decrease in smell recognition that accompanies Alzheimer's corresponds approximately to that experienced by healthy subjects over the course of 30 years. All other factors being equal, if an average healthy person recognized some olfactory stimulus correctly with a probability of 0.5, then the average person with AD would have a probability of about 0.2 of making a correct identification. Finally, women showed olfactory recognition abilities similar to those of men, in the age range studied, and we found that estrogen treatment did not affect healthy women's olfactory recognition abilities.

Some idiosyncrasies were found in subjects' incorrect stimulus identifications, the most important perhaps being that Alzheimer's patients often identified the UPSIT "smoke" sample as a "peach" smell – thus raising concerns about their safety when alone. Also, order effects were observed in subjects' incorrect choices. Together these findings suggest that healthy and AD subjects use different ways to respond to the UPSIT task. Although such group specific response strategies would almost certainly go unnoticed in a classical framework, Rasch scaling provides the tools to assess the nature of these differences. Research is currently in progress to understand the differential distractor responses more fully. In particular, given the theoretical work on Rasch fit indices [12], it appears feasible to construct disease specific person fit indices (i.e., statistics that quantify people's incorrect responses). A possible outcome of this research is that the UPSIT can be augmented with one or more disease specific classification indices. If successful, more powerful tools would become available to aid in the identification of the subtle changes in olfactory performance that are indicative of the early stages of Alzheimer's disease.

Acknowledgements

This research was supported in part by grants from the NIH (PHS# 6-22369), State of Illinois Alzheimer Disease Assistance Act (IDPH# 623409), and the SIU Foundation.

References

[1] T.G. Bond and C.M. Fox, *Applying the Rasch model: Fundamental measurement in the human sciences,* Lawrence Erlbaum, Mahwah, NJ, 2001.

[2] H.E. Brogden, The Rasch model, the law of comparative judgment and additive conjoint measurement, *Psychometrika* **42** (1977), 631–634.

[3] A. Burns, Might olfactory dysfunction be a marker of early Alzheimer's disease? *Lancet* **355** (2000), 84–85.

[4] M.H. Daniel, Behind the scenes: Using new measurement methods on DAS and KAIT, in: *The new rules of measurement: What every psychologist and educator should know,* S.E. Embretson and S.L. Hershberger, eds, Lawrence Erlbaum, Mahwah, NJ, 1999, pp. 37–64.

[5] D.P. Devanand, K.S. Michaels-Marston, X. Liu, G.H. Pelton, M. Padilla, K. Marder, K. Bell, Y. Stern and R. Mayeux, Olfactory deficits in patients with mild cognitive impairment predict Alzheimer's disease at follow-up, *American Journal of Psychiatry* **157** (2000), 1399–1405.

[6] R.L. Doty, P. Shaman and M. Dann, Development of the University of Pennsylvania Smell Identification Test: A standardized microencapsulated test of olfactory function (monograph), *Physiol Behav* **32** (1984), 489–502.

[7] S.E. Embretson and S.L. Hershberger, *The new rules of measurement,* Lawrence Erlbaum Associates, Mahwah, NJ, 1999.

[8] A.B. Graves, J.D. Bowen, L. Rajaram, W.C. McCormick, S.M. McCurry, G.D. Schellenberg and E.B. Larson, Impaired olfaction as a marker for cognitive decline, *Interaction with apolipoprotein E ε4 status* **53** (1999), 1480–1487.

[9] J. Hattie, Methodology review: Assessing unidimensionality of tests and items, *Applied Psychological Measurement* **9** (1985), 139–164.

[10] R.I. Henkin, Olfactory dysfunction in Alzheimer's disease, *Lancet* **355** (2000), 1014.

[11] L.F. Hughes, R.G. Struble and C.L. Shaffer, Olfaction in elderly and Alzheimer patients: differing feedback conditions, *J Alzheimer's Disease* **3** (2001), 367–375.

[12] G. Karabatsos, A critique of Rasch residual fit statistics, *Journal of Applied Measurement* **1**(2) (2000), 152–176.

[13] R. Lange, H.J. Irwin and J. Houran, Top-Down Purification of Tobacyk's Revised Paranormal Belief Scale, *Personality and Individual Differences* **29** (2000), 131–156.

[14] R. Lange, M.A. Thalbourne, J. Houran and L. Storm, The revised transliminality scale: Reliability and validity data from a Rasch Top-Down Purification procedure, *Consciousness and Cognition* **9** (2000), 591–617.

[15] J.M. Linacre, *Many-facet Rasch measurement,* MESA Press, Chicago, 1989.

[16] J.M. Linacre, *A user's guide to Facets Rasch measurement computer program*, MESA Press, Chicago, 2001.

[17] F.M. Lord, *Application of item response theory to practical problems Hillsdale,* Lawrence Erlbaum Associates, NJ, 1980.

[18] R.I. Mesholam, P.J. Moberg, R.N. Mahr and R.L. Doty, Olfaction in Neurodegenerative Disease: A meta-analysis of Olfactory Functioning in Alzheimer's and Parkinson's Diseases, *Archives of Neurology* **55** (1998), 84–90.

[19] R.J. Mislevy and R.D. Bock, *BILOG 3: Item analysis and test scoring with binary logistic models,* Scientific Sofware Inc., Chicago, IL, 1990.

[20] M.J. Moore, C.W. Zhu and E.C. Clipp, Informal costs of dementia care: Estimates from the National Longitudinal Caregiver Study, *Journal of Gerontology: Social Sciences* **56B**(4) (2001), S219–S228.

[21] R. Perline, B.D. Wright and H. Wainer, The Rasch model as additive conjoint measurement, *Applied Psychological Measurement* **3** (1979), 237–255.

[22] K. Perkins, B.D. Wright and J.K. Dorsey, Using Rasch measurement with medical data, in: *Rasch measurement in health sciences,* Mesa Press, Chicago, (in press).

[23] G. Rasch, *Probabilistic models for some intelligence and attainment tests,* MESA Press, Chicago, IL, 1960.

[24] W.F. Stout, A nonparametric approach for assessing latent trait dimensionality, *Psychometrika* **55** (1987), 293–326.

[25] The Lewin Group report, Alzheimer's Association, 919 North Michigan Avenue, Suite 1100, Chicago, Illinois 60611-1676, April 3, 2001.

[26] B.D. Wright, Fundamental measurement for psychology, in: *The new rules of measurement,* S.E. Embretson and S.L. Hershberger, eds, Lawrence Erlbaum Associates, Mahwah, NJ, 1999, pp. 65–104.

[27] B.D. Wright and M.H. Stone, *Best Test Design,* MESA Press, Chicago, IL, 1979.

[28] M.L. Wu, R.J. Adams and M.R. Wilson, ConQuest: Generalized item response modeling software Australian Council for Educational Research (ACER), 1998.