Chem. Senses 25: 493–499, 2000

Psychometrics of Odor Quality Discrimination: Method for Threshold Determination

Mats J. Olsson and William S. Cain[1]

Department of Psychology, Uppsala University, Box 1225, S-751 42 Uppsala, Sweden and [1]Chemosensory Perception Laboratory, University of California at San Diego, La Jolla, CA 92093-0957, USA

Correspondence to be sent to: Mats J. Olsson, Department of Psychology, Uppsala University, S-751 42 Uppsala, Sweden. e-mail: mats.olsson@psyk.uu.se

Abstract

There is no natural physical continuum for odor quality along which an odor quality discrimination (OQD) threshold can be measured. In an attempt to overcome this problem, the substitution–reciprocity (SURE) method defines a framework for the measurement of an OQD threshold. More specifically, it (i) defines a threshold concept for OQD, including the quantification of qualitative change of the stimulus, and (ii) suggests how to avoid perceived intensity as an unwanted cue for discrimination. In doing this, the psychometric properties of odor quality in the case of eugenol and citral are investigated using both discrimination (experiment 1) and scaling (experiment 2). Based on discriminatory responses, a change of approximately one-third in stimulus content was needed to reach the OQD threshold for eugenol and citral.

Introduction

There is no standardized psychometric method for the measurement of odor quality discrimination (OQD) thresholds available. The main reason is that there is no natural physical continuum for odor quality, such as sound frequency in hearing. Especially in clinical research, OQD is commonly measured as confusability of odors using a same–different or oddity procedure (Wright, 1987; Martinez *et al.*, 1993; de Wijk and Cain, 1994; Savic *et al.*, 1997; Laska and Teubner, 1999a,b) [for a review see (Wise *et al.*, 2001)]. Although these 'confusion tests' may serve the intended purpose and are easy to administer, there is a reason why a psychometric test of OQD would be preferred. Studies have shown that hit and correct rejection rates in confusion tests of quality discrimination are high, often 90% (de Wijk and Cain, 1994), unless very similar odorants are compared (Laska and Teubner, 1999a,b). Performance levels of confusion tests may therefore be too poor to differentiate effectively between different pairs of odors or between individuals.

There is one main confounding factor in designing a psychometric OQD threshold test: the *perceived intensity* of the odors to be compared. The problem is twofold.

The first problem relates to observations made in odor mixture research. In a study of a binary odor mixture of pyridine and 1-butanol, Olsson (Olsson, 1994, 1998) found that a mixture of two components will yield a mixture quality where the qualitative dominance of the stronger component is proportional to the difference between the squared perceived intensities of the unmixed components. This observation was formalized in a model stating that the probability (P) that a substance A will dominate a mixture of A and B depends on the perceived intensity (R) such that $P(\mathrm{a}) = R_A^2/(R_A^2 + R_B^2)$. In other words, the olfactory system seems to accentuate the difference in intensity between two odors when determining a mixture quality. Therefore, an OQD procedure comparing a standard odor with a comparison odor (A), consisting of the standard mixed with preset fractions of a second odor (B), will yield a threshold that is more or less dependent on how well the two odorants A and B are matched for perceived intensity.

The second reason to regard perceived intensity as a confounding factor concerns the fact that intensity itself can provide an unwanted cue to discrimination. If OQD is investigated using an additive procedure, where a standard is compared with itself with variable amounts of another odor added to it (Bende and Nordin, 1997), the standard and comparison stimuli can have quite different perceived intensities depending on how much of the adulterant odor has to be added before a criterion change in quality has been reached. A method for measuring OQD thresholds should therefore take this into account.

The general aim of this paper is to propose and test a method that defines an OQD threshold concept for pairs of odorants and that counteracts confounding effects of

perceived intensity. The two elements of the method are *substitution* and *reciprocity*. The substitution–reciprocity (SURE) method defines a threshold concept for OQD and could be combined with other procedural choices pertaining to whether the measurement should be quick or reliable. The first element of the SURE method is that it varies the quality of odors to be discriminated through substitution of odorants rather than by addition of one odorant to another. This means that instead of adding a fraction of an adulterant to a standard stimulus in order to form a comparison stimulus, the same fraction is first removed from the standard stimulus quantity before they are mixed. The reason is to keep the standard and comparison stimuli at about the same level of perceived intensity and to cause a difference only in odor quality. It should be noted at this point that we are not interested in the assessment of quality in an absolute sense, but rather in the measurement of qualitative change, which is operationlized as a change in discriminability. According to results from mixture research (Laska and Hudson, 1991; Cain *et al.*, 1994; Wise and Cain, 2000) substitution will lead to roughly equal intensities. The validity of this assumption will be tested in the second experiment of the current study. That odor quality will vary monotonically from percept A to percept B as the stimulus is gradually changed from A to B is well supported by several studies (Ekman *et al.*, 1964; Moskowitz, 1976; Laing *et al.*, 1984; Olsson, 1994; Wise and Cain, 2000) and is also implied by the results in experiment 2. That is, as the physical stimulus varies from containing largely A to largely B, the relative frequency of responses associated with percepts A and B will reflect that change. Whether this qualitative change follows a straight line between point A and B in a perceptual space for odor quality or not is another question that depends on the perceptual processing specific to the olfactory systems and, possibly, the mixing technique. That this may be the case is, however, indicated by a few studies mapping odors in perceptual spaces with different techniques for assessing their qualitative proximity (Ekman *et al.*, 1964; Moskowitz, 1976; Wise and Cain, 2000).

As the second element of the SURE method (reciprocity), it is proposed that the threshold measurement should employ both the odors to be discriminated, A and B, as standards *and* as adulterant stimuli. This means that two *separate* threshold values will be assessed and averaged into a *combined* threshold value. This has two advantages. Thereby the method returns a single value of discriminability associated with the stimulus continuum. Second, the assumption is made that the effects on measures of OQD thresholds caused by possible differences in intensity between odor A and B will cancel out to some degree. That is, if standard stimulus A is stronger than adulterant B, substitution of A with B would yield larger separate threshold values than if stimulus A and B were matched for perceived intensity. Consequently, if B was the comparison stimulus and A the adulterant, the threshold value would then be smaller. Averaging thresholds in this way will contribute to the cancellation of threshold biases that are due to stimulus intensity differences between the two odorants for which the OQD threshold will be assessed. Ideally, perceived intensities should be perfectly matched, but since even minor deviations from a perfect match may affect the quality of the comparison stimulus substantially (Olsson, 1994, 1998), the averaging procedure could increase measurement reliability.

To conclude, through this procedure an OQD threshold for any two odorants could be defined and measured as the average of two fractions in which one odorant has been substituted by the other to reach a criterion level of discriminatory response.

Materials and methods

Experiment 1

Aim

This experiment applied the SURE method to determine the OQD threshold for eugenol and citral. In order to investigate the nature of OQD, individual psychometric functions were assessed and compared. Different ways to calculate the OQD threshold will be discussed in the light of the data.

Participants

Three females (P1, P3 and P6) and three males (P2, P4 and P5) participated in the two experiments. Three were members of staff (P3, P4 and P5) and the other three were students coming to our laboratory for course credits (P1, P2 and P6). They ranged in age from 22 to 53 years. Participants had functional senses of smell according to tests of absolute thresholds for eugenol and citral and self-reports. One participant (P6, a 24-year-old female) reported that she became congested during the repeated smelling and her data were therefore excluded from the general data analysis.

Stimuli

Two standards of eugenol and citral were mixed with mineral oil to concentrations of 0.09 and 0.19% (v/v), respectively. These had been determined to be of about equal perceived intensity in a pilot experiment. Nine liquid phase mixtures of these base concentrations were prepared, ranging from a mixture of 90% eugenol standard and 10% citral standard to a mixture of 10% eugenol and 90% citral (via 80/20, 70/30 and so on). Altogether, 11 unique stimuli were used.

Odors were presented from squeeze bottles (polyethylene, 270 ml volume with 30 ml liquid) with flip-up spouts. Four bottles of each standard and two bottles of each comparison odor were used in order to promote saturation of the head space in the bottle.

Procedure

Participants discriminated among odors in an ABX design,

i.e. they had to decide which of three stimuli was different from the other two. First they smelled two bottles, of which one was a standard (eugenol or citral) and the other was one of the 10 other stimuli (including the other standard). After smelling those two, the third odor presented was either identical to the first or the second odor and the participants' task was to decide which. Every unique pairwise combination (i.e. $2 \times 10 = 20$) was presented 24 times to each subject at a rate of two comparisons per minute. Altogether, each participant made 480 comparisons distributed over six sessions on separate days. A session took about 2 h with a 5 min break. Participants could also take a break if needed at any time. When given a triad of bottles, a participant placed the spout just beneath his/her nostrils, squeezed the bottles at a pace and manner deemed most suitable and inhaled the odor dirhinally.

Experiment 2

Aim

In this experiment participants were asked to estimate the perceived intensity, both overall and component specific, of the same stimuli as in experiment 1. The aims were: (i) to validate, across methods, the measures of discriminability found in the previous experiment; (ii) to test the assumption that the substitution procedure produces a series of comparison mixtures that are of comparable perceived intensity to the standards.

Procedure

The participants, stimuli and concentrations used in this experiment were identical to those of experiment 1 with the exceptions described below.

Eleven test stimuli (the two standards and nine mixtures) were compared with each standard 16 times. Note that this time a standard was also compared with itself. The total of 352 trials per person were presented in four 2 h sessions, on four separate days, comprising 88 trials each. One 5 min break was scheduled but participants could ask for a break at any time if needed. In each trial, the participant smelled two odorants: one of the two standards and then a comparison stimulus that could be any of the 11 different test stimuli. Two of the participants (P1 and P3) could take part in only two of the four sessions. Consequently, they estimated each unique combination of standard and comparison stimuli only eight times.

On a trial, two judgements were given by the method of magnitude estimation (Gescheider, 1997). First, the subject estimated the *overall* perceived intensity of the comparison stimulus in relation to the standard, which had been assigned a value of 100. Second, the responder attended to the perceived quality of the standard in order to judge the perceived intensity of that specific quality in the comparison odor. For the estimation of this *component-specific* intensity, the modulus was again set to 100.

Results and discussion

Experiment 1

Measurement of OQD threshold

For each unique combination, proportions of correct discriminations [P(c)] were calculated for each of the five participants. P(c) values for the two standards (eugenol and citral) were plotted as a function of proportion of citral in the mixture (Figure 1). The individual and group results indicated that it took proportionally less citral to adulterate eugenol than vice versa (Figure 1). Only participant P2 found it otherwise.

The OQD thresholds were determined by first pooling the two separate functions generated for the two standards (Figure 2). This was done by averaging the two P(c) values for each percentage of adulterant in the mixture. Second, a logistic function was fitted to the pooled data from which a combined threshold value (T_C) could be read for P(c) = 0.75, half way between chance and perfect performance. Assuming that the gaseous concentrations were proportional to the liquid concentrations (Raoult's law), the OQD thresholds (T_C) among the five participants ranged from 0.25 to 0.42, using a criterion of 75% correct. The group data yielded a threshold of 0.34. This means that by conventional analysis a 34% change in content was just noticeable.

Classical threshold theory is often identified with the so-called ϕ–γ hypothesis. This means that the probability of a response (ϕ) as a function of stimulus change (γ) should have the ogival form of the cumulative normal distribution. An alternative hypothesis is the ϕ–logγ hypothesis that states this specific relationship to be true only when stimulus change is given as logarithms (Gescheider, 1997, p. 80). Before proportions of correct discrimination [P(c)] were transformed into Z scores, the chance performance level (0.5) was subtracted to form a new measure [$P(c_0)$] that varied from approximately 0 to 1: $P(c_0) = [P(c) - 0.5]/[1 - 0.5]$. After transformation of proportions of correct discrimination to Z scores for the present data, the psychometric function is better described by a straight line when the stimulus proportions were given as logarithms ($r^2 = 0.96$; Figure 3) than when given in linear terms ($r^2 = 0.93$). In other words, the OQD function lends some support to the ϕ–logγ hypothesis. An alternative procedure to estimate the OQD threshold for this data set is to use the regression equation given in Figure 3, which yielded 0.29. Another possible virtue of Z scoring the data is that the slope of the regression line is a simple measure of transition sharpness, i.e. the stimulus range necessary to go from chance to perfect performance, which could possibly reflect the degree to which categorical perception is at work.

Confusion data

As mentioned earlier, the sense of smell is not perfect in discriminating between even dissimilar odorants, at least to judge from confusion tests and mixture research. Probably,

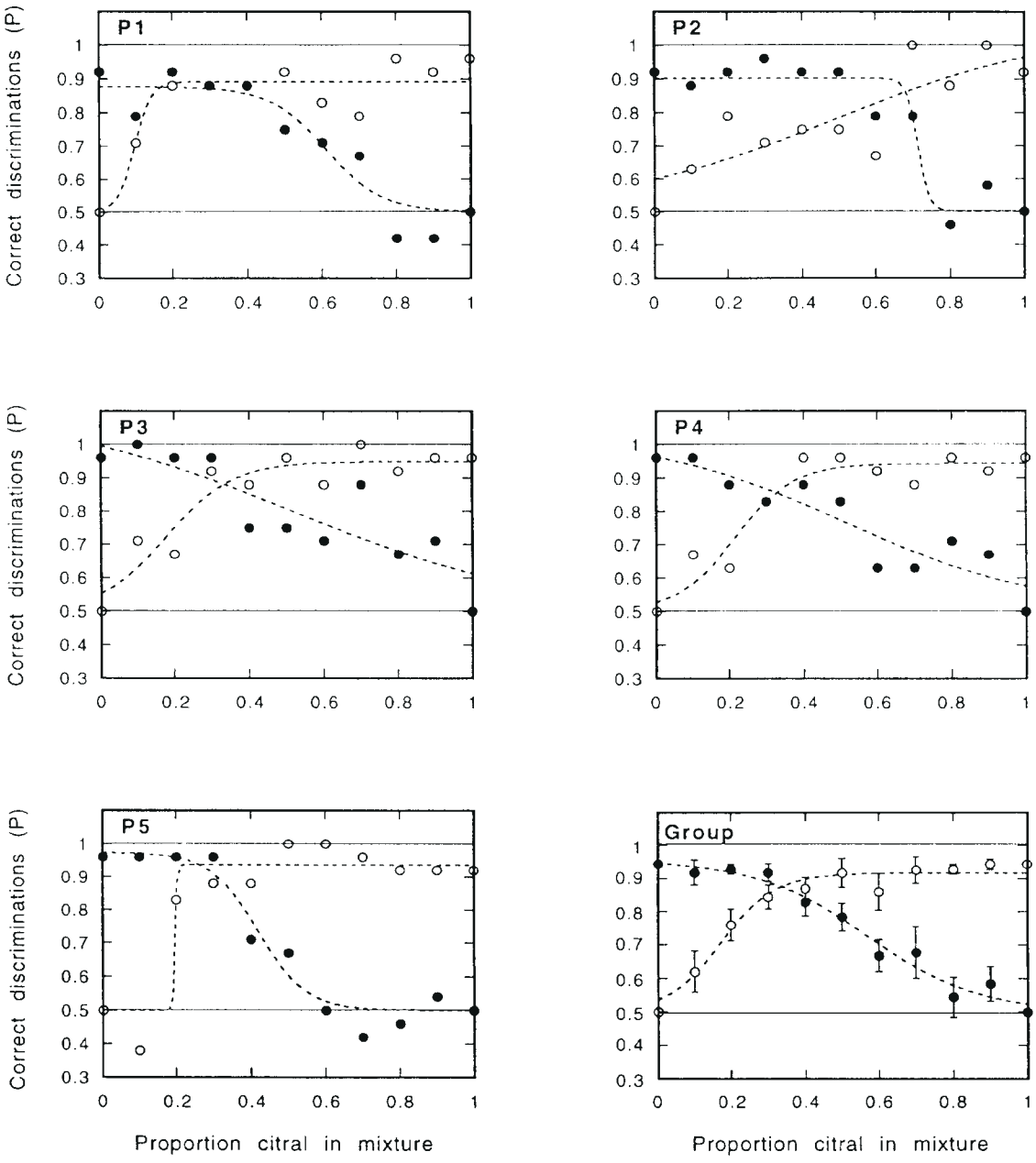

Figure 1 Psychometric functions for quality discrimination for eugenol, citral and mixtures of them shown for individual participants and for the group (±SE). Open circles denote proportions of correct discrimination between a standard of pure eugenol and a mixture where citral was substituted for eugenol. Filled circles denote the reversed case where citral was the standard and eugenol the substitute. Theoretical data points [$P(c) = 0.5$] were added for the two theoretical cases were the standard would have been compared with itself. This was done in order to stabilize the logistic function fitted to the data: $y = 0.5+(c-0.5) \times \{\exp(a + bx)/[1 + \exp(a + bx)]\}$.

few people would be of the opinion that lemon (citral) and cloves (eugenol) are confusable. Yet, confusions do occur. The proportion of times (out of 48) that pure concentrations of eugenol and citral were confused varied between 4 and 8% among participants (Figure 1). Theoretically, these relatively small confusion rates could depend on other factors than discriminablility, such as adaptation, attention and memory. For the current data there is a positive correlation between the individual OQD thresholds and confusion rates, but this did not reach statistical reliability for this small number of participants ($r = 0.61$, df = 4, P not significant).

Experiment 2

Estimates of overall and component-specific odor intensities were averaged for each unique comparison across the 16 repetitions. To investigate whether the procedures in the previous and current experiment gave a comparable

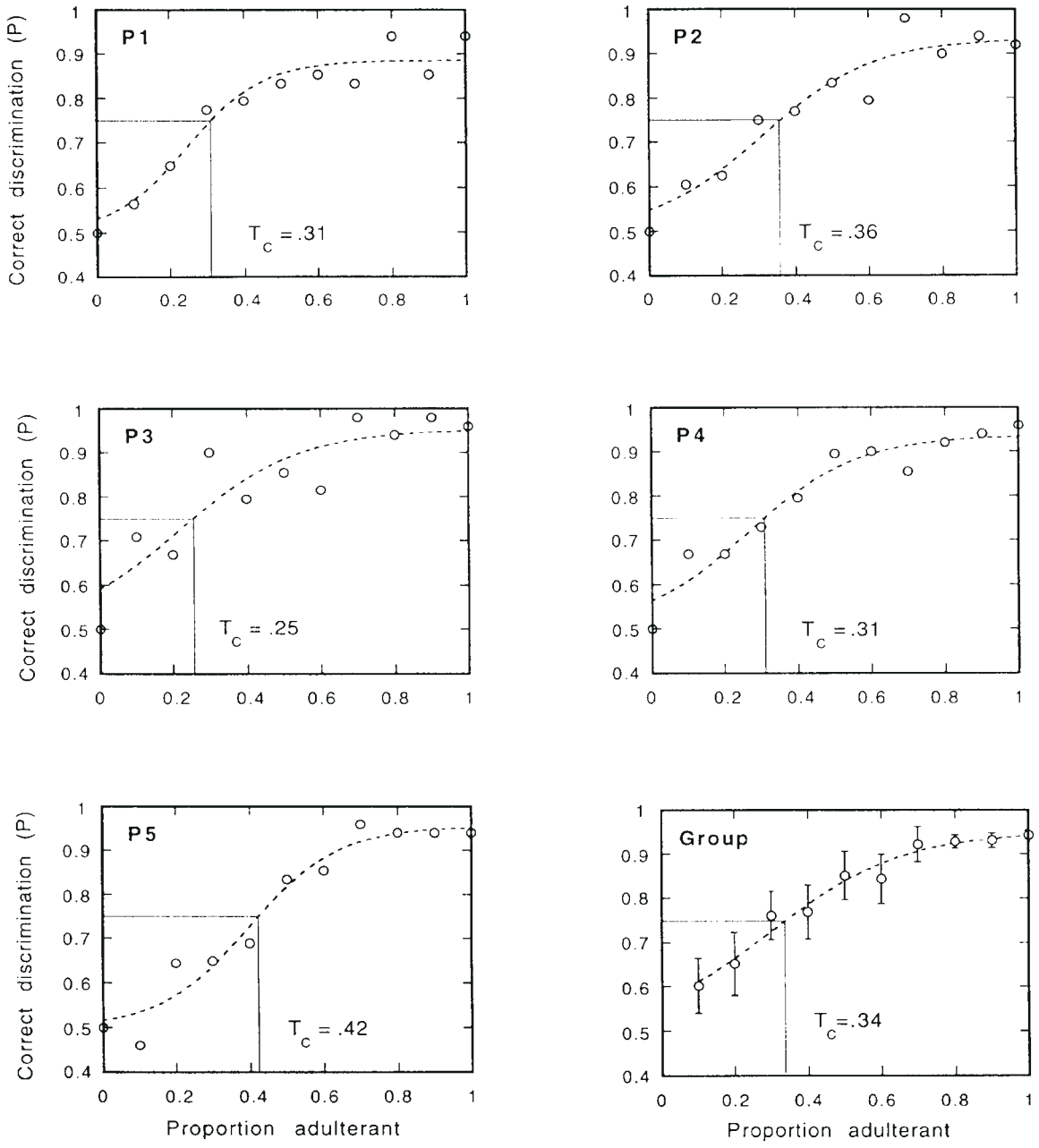

Figure 2 The psychometric functions for the combined functions of Figure 1. The combined thresholds (T_C) are given for individual participants and the group. A theoretical data point was added for the case were the standard would have been compared with itself [$P(c) = 0.5$].

outcome with respect to change of quality as the stimuli varied from citral to eugenol, estimated component-specific intensity as a fraction of overall intensity was related to stimulus content (Figure 4). The overall picture of how perception changes as the content of the stimulus changes was roughly the same between the two methods. The points on the abscissa where the two qualities are subjectively equally prominent, i.e. where the two functions cross, agree fairly well between the methods (0.34 in Figure 1 and 0.30 in Figure 4). In other words, both measures indicate a 'dominance' of citral over eugenol since citral is easier to detect when occuring as a contaminant of eugenol compared with the reversed case. However, this dominance does not necessarily reflect a characteristic of citral, but rather the fact that the concentration of citral chosen for this experiment turned out to be subjectively stronger than the eugenol standard. Citral was judged to be stronger than eugenol by all participants, averaging ~17.4% stronger in subjective units (range 10–39%).

In order to monitor the overall perceived intensity as a function of stimulus content, Figure 5 was plotted. First, it is obvious that the standard citral stimulus was stronger

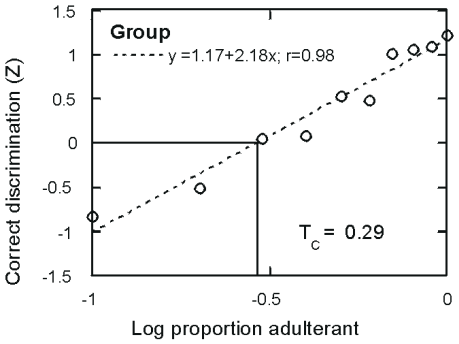

Figure 3 The psychometric group function of Figure 2 where $P(c)$ is transformed to Z scores. The combined group threshold (T_C) could here be calculated through the linear regression equation.

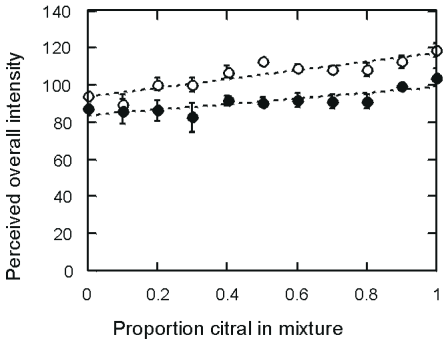

Figure 5 Perceived overall intensity as a function of the proportion of citral in the mixture (comparison stimulus). Open symbols denote the case where eugenol was the standard and filled symbols where citral was the standard (modulus = 100 in each case).

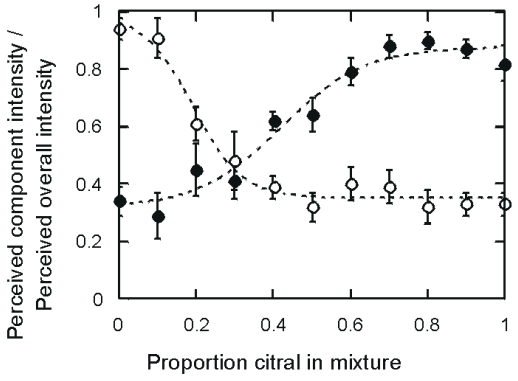

Figure 4 Perceived intensity of the citral and eugenol components, divided by overall mixture intensity, as a function of proportion citral in the mixture (comparison stimulus). Open symbols denote the case where eugenol was the standard, and hence the quality to be estimated, and filled symbols denote the case where citral was the standard.

than the standard eugenol stimulus, wherefore the regression line in Figure 5 has a slope. However, more importantly, the overall intensity estimates of the different mixtures fit well to the linear regression line, which indicates that the substitution procedure itself did not produce any systematic changes in perceived intensity of the stimulus over and above that produced by mismatched standards. In other words, the results in Figure 5 support the assumption that substitution counteracts the problem of perceived intensity being a cue to odor quality discrimination which could be present in methods employing an additive procedure.

General discussion

In summary, there has been no standardized psychometric method to measure OQD thresholds to yield an estimate of a threshold value. In this paper elements of such a method are proposed and tested. The substitution–reciprocity (SURE) method defines a threshold value that is an average fraction by which one odorant has to be substituted with another to reach a criterion level of discrimination, i.e. the method returns a single measure of discriminability for a physical continuum ranging from A to B. The method counteracts perceived intensity as a confounding factor through the use of a substitution procedure which permits variation of the quality of an odor stimulus without making perceived intensity a cue for a discriminative response. Moreover, the measure of discrimination is reciprocal in the sense that the measure of the OQD threshold is a result of two separate psychometric functions involving two different standards but otherwise the same comparison stimuli. This reciprocity will counteract threshold biases that are due to an intensity mismatch between comparison stimuli.

The current data for eugenol and citral suggest that a change of approximately one third in stimulus content is needed to reach threshold performance. A second experiment confirmed the assumption of the SURE method that the substitution procedure did not affect overall odor intensity of the comparison stimuli (mixtures) as the quality of these are varied, thereby counteracting an irrelevant cue for discrimination.

In comparing the discrimination data (experiment 1) and scaling data (experiment 2), similar pictures of how perceived quality varies with stimulus change emerged. However, one interesting difference is notable. The functions in Figure 2 show that discrimination data resolved smaller contamination of the standard than did scaling data (Figure 4). The poor resolution of qualitative change using scaling could possibly be explained by a conservatism on the participant's behalf preventing him/her from stating that a standard smells only of that particular quality. The fact that discrimination data also failed to show perfect resolution for large stimulus differences suggests that confusion tests tap asymptotic performance. How this level of asymptotic performance is related to the psychometric threshold is certainly an interesting topic for future research.

On the issue of measuring small changes in quality, it should be noted that the threshold determinations in Figures 2 and 3 using the 75% correct criterion were based on

comparisons that mostly surpassed that level of discrimination. Therefore, it may be advisable to consider geometric step sizes instead of arithmetic ones in varying stimulus content. This refinement of the SURE method would have made the interpolation of the threshold value more reliable, at least in this case. Another refinement concerns the best method for pooling the two separate psychometric functions to form a combined threshold value. This is not known and should be the subject of further research. The criterion for a suitable averaging procedure is that the combined threshold when standards do not match should approximate the combined threshold when standards are perfectly matched.

One line of research for the future concerns how qualitative discriminability varies with properties of the stimulus, such as intensity and quality. Another line of research concerns the types of problems in theoretical psychophysics to which the proposed method of measuring OQD thresholds could be applied. One such problem is the measurement of similarity. As a first step, it would be of interest to validate the OQD threshold against other measures of similarity, such as direct judgements of similarity (Moskowitz, 1974) or inferred measures of similarity based on reaction time (Wise and Cain, 2000).

Yet another psychophysical problem that could benefit from psychometric measurements of OQD concerns how to measure the masking potency of odors. Since there is no standardized framework for such measurements, there are also very few comparable results on masking potency published in the literature. Hence, formal knowledge regarding which odors are good maskers are almost non-existent. Possibly, the difference between separate thresholds for intensity-matched standards could provide an index of masking potency, thereby providing comparable measures to promote knowledge in this field.

Acknowledgements

The authors are grateful to Steven Nordin and Paul Wise for comments on a previous version of this manuscript. This study was made possible by grant RO1 DC00284 from the US National Institute on Deafness and Other Communication Disorders.

References

Bende, M. and **Nordin, S.** (1997) *Perceptual learning in olfaction: professional wine tasters versus controls.* Physiol. Behav., 62, 1065–1070.

Cain, W.S., Schiet, F.T., Olsson, M.J. and **de Wijk, R.A.** (1995) *Comparison of models of odor interaction*. Chem. Senses, 20, 625–637.

de Wijk, R.A. and **Cain, W.S.** (1994) *Odor quality: discrimination versus free and cued identification.* Percept. Psychophys., 56, 12–18.

Ekman, G., Engen, T., Künnapas, T. and **Lindman, R.** (1964) *A quantitative principle of qualitative similarity.* J. Exp. Psychol., 68, 530–536.

Gescheider, G.A. (1997) Psychophysics: The Fundamentals, 3rd edn. Lawrence Erlbaum Associates, London.

Gregson, R.A.M. (1980) *A model of paradoxical odour mixture perception*. Chem. Senses Flav., 5, 257–269.

Laing, D.G., Panhuber, H., Willcox, M.E. and **Pittman, E.A.** (1984) *Quality and intensity of binary odor mixtures.* Physiol. Behav., 33, 309–319.

Laska, M. and **Hudson, R.** (1991) *A comparison of the detection thresholds of odour mixtures and their components*. Chem. Senses, 16, 651–662.

Laska, M. and **Teubner, P.** (1999a) *Olfactory discrimination ability for homologous series of aliphatic alcohols and aldehydes.* Chem. Senses, 24, 161–170.

Laska, M. and **Teubner, P.** (1999b) *Olfactory discrimination ability of human subjects for ten pairs of enantiomers.* Chem. Senses, 24, 263–270.

Martinez, B.A., Cain, W.S., de Wijk, R., Spencer, D.D., Novelly, R.A. and **Sass, K.J.** (1993) *Olfactory functioning before and after temporal lobe resection for intractable seizures.* Neuropsychology, 3, 351–363.

Moskowitx, H.R. (1976) *Multidimensional scaling of odorants and mixtures.* Lebensmittelwiss. Technol., 9, 232–238.

Olsson, M.J. (1994) *An interaction model for odor quality and intensity.* Percept. Psychophys., 55, 363–372.

Olsson, M.J. (1998) *An integrated model of intensity and quality of odor mixtures.* Ann. NY Acad. Sci., 855, 837–840.

Savic, I., Bookheimer, S.Y., Fried, I. and **Engel, J.** (1997) *Olfactory bedside test: a simple approach to identify temporo-orbitofrontal dysfunction.* Arch. Neurol., 54, 162–168.

Wise, P.M. and **Cain, W.S.** (2000) *Latency and accuracy of discriminations of odor quality between binary mixtures and their components*. Chem. Senses, 25, 247–265.

Wise, P.M., Olsson, M.J. and **Cain, W.S.** (2001) *Quantification of odor quality*. Chem. Senses, 26, in press.

Wright, H.N. (1987) *Characterization of olfactory dysfunction*. Arch. Otolaryngol. Head Neck Surg., 113, 163–168.

Accepted February 21, 2000