Latency and Accuracy of Discriminations of Odor Quality between Binary Mixtures and their Components

Paul M. Wise and William S. Cain

Chemosensory Perception Laboratory, Department of Surgery (Otolaryngology), University of California, San Diego, La Jolla, CA 92093-0957, USA

Correspondence to be sent to: Paul M. Wise, Chemosensory Perception Laboratory, Department of Surgery (Otolaryngology), Mail Code 0957, University of California, San Diego, La Jolla, CA 92093-0957, USA. e-mail: pwise@ucsd.edu

Abstract

Subjects made timed, same–different discriminations of odor quality, with the following principal findings: (i) latency reflected accuracy, with difficult discriminations, i.e. those between 50–50 mixtures and their components, requiring more time than less difficult discriminations, i.e. those between unmixed chemicals. This finding demonstrated the face validity of latency as a measure of qualitative similarity. (ii) Latency provided better resolution among pairs of odors than did errors of discrimination. This finding demonstrated the utility of collecting response times. (iii) Latency-based similarities among odors tested previously predicted similarities among pairs not yet tested. This finding demonstrated internal/predictive validity. (iv) A signal detection model assuming a differencing strategy best described the pattern of errors. Subjects appeared to make relative judgements regarding quality. (v) Finally, latency-based similarities between mixtures and their components demonstrated additivity. This finding suggested that binary mixtures fall on straight lines connecting their components in 'odor-space'.

Introduction

Discrimination of odor quality, instantiated in a variety of paradigms, measures subjects' ability to resolve differences in kind among odors. This paper examines the same–different paradigm, where subjects receive some pairs of odors that differ in quality and some that do not. Since frequency of discriminative errors will increase as stimuli become more similar, discriminability constitutes a performance-based (objective) operational definition of qualitative similarity.

Unfortunately, once odors differ beyond a certain extent, performance may approach an asymptote. This has abridged the use of discrimination to studies of fairly similar-smelling molecules (Laska and Freyer, 1997; Laska and Teubner, 1999; Laska *et al.*, 1999). A new dependent variable may allow discrimination to map qualitative differences over a wider range of molecular parameters. A large literature suggests that the time needed to compare two stimuli varies inversely with the distance between them on the judged dimension, even after stimuli become perfectly discriminable (Woodworth and Schlosberg, 1954; Welford, 1960; Vickers, 1980; Luce, 1986). Could latency extend the range of differences discrimination can resolve?

Basic questions regarding latency to discriminate

Latency to discriminate reflects similarity in other sensory modalities, but it also reflects emphasis on speed versus accuracy and response-bias, i.e. a bias toward responding 'different' (Pike, 1971; Ratcliff and Hacker, 1981; Proctor and Weeks, 1989; Ratcliff and Rouder, 1998). From the standpoint of measurement, these problems can be solved with a sustained effort. Prior to this effort, several basic issues require attention.

The work should establish that: (i) difficult judgements, or discriminations between similar odors, take longer than less difficult judgements, or discriminations between less similar odors. One can manipulate quality by, for example, diluting chemicals with one another. Binary mixtures should smell more similar to their unmixed components than the unmixed components smell to each other. Discriminations between mixtures and components should accordingly result in more errors and require more time than discriminations between unmixed components. The first of two studies will examine this issue. (ii) Latency to discriminate provides better resolution among odor-pairs than errors of discrimination. The first study will examine this issue. (iii) Discriminabilities among odors tested previously should predict discriminabilities among pairs not yet tested. The second study will examine this issue.

Signal detection models of same–different discriminations

The theory of signal detection (TSD) provides an index of discriminability (d') that is not sensitive to bias and allows metric comparisons between differences in sensation (Swets *et al.*, 1961). Signal detection models of the same–different task assume that a normal, unidimensional distribution can represent the sensations a stimulus produces over many trials, and that distributions of sensation associated with the two stimuli presented during a trial have equal variance. Both assumptions have proven reasonable (Laming, 1986). However, signal detection models of the same–different task also require knowledge of the decision-strategy that subjects apply (Macmillan and Creelman, 1991).

Subjects might make absolute judgements, e.g. that the first odorant presented during a trial is more likely 'apple' and the second is more likely 'orange'. Some investigators have labeled this the *independent-observation* (Macmillan and Creelman, 1991) strategy. Alternatively, subjects might make only relative judgements, e.g. that two stimuli smell different enough to warrant a response of 'different'. Investigators have labeled this the *differencing* strategy (Macmillan and Creelman, 1991; Irwin and Francis, 1995). Which strategy will subjects apply in same–different discriminations of odor quality? No data exist. Yet, since the two strategies lead to different levels of performance as measured by percent correct, computation of d', which relies on percent correct, requires information regarding strategy.

ROC (receiver operating characteristic) curves reflect strategy. They show the proportion of 'hits', or nominally different pairs that subjects call 'different', versus the proportion of 'false alarms', or nominally same pairs that subjects call 'different', at each of a number of criteria (Macmillan and Creelman, 1991). Differencing leads to an ROC curve asymmetric with respect to the negative diagonal of the unit square in which ROC curves are plotted (Macmillan *et al.*, 1977; Irwin *et al.*, 1993), whereas independent observations lead to an ROC curve symmetric with respect to the negative diagonal (Francis and Irwin, 1995, 1997; Irwin and Hautus, 1997). The second study will determine which model provides the best fit for same–different ROC curves based on rated confidence.

The perceived quality of binary mixtures

Although this report deals primarily with measurement, it provides information regarding the qualitative relationship between 50–50 mixtures and their components. Results from a variety of methods suggest the perceived qualities of binary mixtures fall between the qualities of their components on at least one dimension (Laing and Willcox, 1983; Laing *et al.*, 1984; Lawless, 1989; MacRae *et al.*, 1990, 1992; Olsson, 1994), but they may differ on at least one other dimension (Olsson, 1994). Do binary mixtures lie along a straight line connecting their components in 'odor-space', or do they lie off that line? The first study will explore this issue by examining the discriminabilities between 50–50 mixtures and their components.

Study 1

Methods

Subjects

Six normosmic, non-smoking females (aged 19–31 years) participated in intensive testing. Two were laboratory assistants and four were paid volunteers.

Apparatus

The main components of the apparatus included: (i) two Lucite bottle-compressors, each of which held two 270 ml plastic bottles, to deliver stimuli; (ii) the program PsyScope, run on a microcomputer, to guide experimenters through the details of trials and to register subjects' responses (Cohen *et al.*, 1993); and (iii) a 'button box' (Research Methods, Pittsburgh, PA) to register responses. Each bottle in a compressor led to one nostril. The second of the two compressors, in the order of use in a trial, included a microswitch that activated a timer in the button box when the experimenter pressed a lever to expel vapor. A button-press stopped the timer, which measured latency of response to within 1 ms.

Teflon tubing of 1/16″ inner diameter channeled odor-puffs from the bottles to the subjects. For each compressor, one tube connected the left bottle to the left nostril and another tube connected the right bottle to the right nostril. The tubes fitted snugly into flip-up spouts of the bottle-closures. The corresponding tubes from the two compressors met at Y-connectors, so the left and right channels from the two compressors shared 6.5 cm of common tubing that ended in flared plastic nose-pieces. Each channel, including the shared portion, measured 55.0 cm long, for a total dead volume of ~1.25 cm^3 per channel. To reduce contamination, all tubing was replaced daily, with the Y-connectors and nose-pieces replaced weekly.

One compressor delivered 50 ± 0.58 and 48.7 ± 0.58 cm^3 (mean ± SD) to the right and left nostrils, respectively. The other compressor delivered 48.4 ± 0.58 and 49.9 ± 0.84 cm^3 to the right and left nostrils, respectively. Average lever-presses lasted 183 ms, with the slowest and quickest of four experimenters differing by only 30 ms. The flow rate at each nose-piece per puff peaked at 2.35 (0.06) l/min. This occurred at the moment the experimenter fully depressed the lever of the compressor. After this, flow decreased at a decelerating rate, and fell to zero at 1050 (120) ms.

Stimuli

The stimuli included both single chemicals (Sigma, St Louis, MO; reagent grade, 97–99.5%) and binary mixtures. Single chemicals, in light mineral oil, included: (i) 1.25% v/v *n*-amyl acetate (banana), (ii) 1% v/v *trans*-anethole (licorice), (iii) 1% v/v benzaldehyde (cherry/bitter almond), (iv) 1% v/v citral (lemon peel), (v) 1% v/v eugenol (oil of clove) and (vi) 1% v/v methyl salicylate (oil of wintergreen). We shall refer to these stimuli as AA, An, B, Cit, Eug and MS, respectively.

A previous study (de Wijk and Cain, 1994) found that these concentrations produced odors of moderate (environmentally realistic) and approximately equal (differences in intensity could serve as discriminative cues, and could therefore become confounded with differences in quality) perceived intensity. Binary mixtures included all unique pairwise combinations of the six odorants. Thus, the stimulus set included six single chemicals and 15 mixtures.

For half the subjects in study 1, Interflo pellets (Chromex Corp., New York, NY) provided the reservoirs of odorant for low-density polypropylene squeeze-bottles. For the other subjects in the first study and all subjects in the second study, 3/4″ diameter felt balls provided the reservoirs for high-density polyethylene squeeze-bottles. Availability of materials prompted the switch. A bottle used to present a single chemical contained four Interflo pellets, each of which held 0.2 ml of the same solution, or two felt balls, each of which held 0.4 ml of the same solution. A bottle used to present a mixture contained two pellets of one solution and two of the other, or one felt ball of one solution and one of the other. Stimulus sets included multiple representatives of identical composition to avoid depleting vapor concentration during a session.

Procedure

Measurement of absolute threshold. To assure that the stimuli lay well and more or less uniformly above threshold, the study included measurements of absolute threshold for the six single odorants. Each subject completed three threshold measurements for each odorant using a two-alternative forced-choice (2-AFC) procedure. Subjects completed all 18 measurements in random order, with an average of three measurements plus two 15 min breaks per 2 h session. One to four days elapsed between sessions.

During a trial, subjects received two bottles, one containing a felt ball with mineral oil and the other containing a felt ball with a dilution of odorant. Subjects indicated which bottle smelled stronger. Subjects sampled bottles by removing the cap, squeezing once and sniffing with the opening directly under both nostrils. At least 45 s separated successive trials. A staircase method was used, moving up one twofold concentration step after an incorrect response and down one step after two consecutive correct responses. Staircases began at the middle of the concentration range (1.0×10^{-2}–7.6×10^{-8}% v/v for Eug, 1.0×10^{-2}–2.4×10^{-6}% v/v for other odorants) for the first measurement and at the previously measured threshold thereafter. Each measurement included six reversals. Threshold was the average of the last five.

Ratings of perceived intensity. In order to ensure that the stimuli, both single chemicals and mixtures, matched one another in intensity for the current subjects, the study included ratings of perceived intensity. Each subject completed three ratings of each of the 21 single and mixed stimuli at three concentrations spanning a 25-fold range (0.25, 1.25 and 6.25% v/v for AA, and 0.20, 1 and 5% v/v for all other odorants). Presentation was dirhinic, using the same apparatus used for discrimination.

Each block of trials included one presentation of each odorant at the three concentrations, for a total of 63 trials. At the start of a block, the experimenter delivered a single puff of a standard (1.25% v/v amyl acetate). Subjects received the standard on demand, or after every fifth trial. A random number generator determined the order of presentation of the 63 trials, with a 1 min interval between trials. Subjects rated the intensity of single puffs of each odor, relative to the standard, by marking a visual analogue scale. The scale consisted of a 156 mm horizontal line-segment, with the term ‘no odor’ on the left. A hash mark at 52 mm indicated the perceived intensity of the standard. A response-packet included only a single scale per page. A full set of ratings required ~105 min, with the three sets completed over a period of 2–3 days.

Discrimination. At the start of a trial the experimenter instructed the subject to place her nostrils over the nose-pieces, and to place her dominant hand on the button-box. The subject placed the nose-pieces far enough into her nostrils to ensure that none of the odor-puff escaped. The subject placed her middle finger between two buttons, with her index and fourth fingers covering the buttons. Three subjects responded ‘different’ with their index fingers, while the other three responded ‘different’ with their fourth fingers.

The experimenter gave a ‘ready’ signal after the subject moved into position, then simultaneously depressed the lever of one compressor and the mouse-button of the computer. The subject sniffed and sought to remember the quality of the odor-puff. The mouse-click began a 6 s inter-stimulus interval (ISI). Pilot work showed that 6 s provided reasonable protection against interference, both physical and through adaptation, between the two odor-puffs. After the ISI, the experimenter pressed the lever of the second compressor. No warning preceded the second odor-puff. Since subjects had completed between 100 and 200 practice trials before collection of data began, they knew the time-course of trials. The second lever-press activated the button-box timer (see Apparatus). The timer stopped when the subject indicated, by pressing one of two buttons, whether or not the two odor-puffs differed in quality.

Instructions regarding speed–accuracy trade-off followed: ‘We will time your responses, but please do not sacrifice accuracy for speed. We do urge you to respond as quickly as you can after you reach a conclusion, but let me stress again that we want your best answer, however long it takes.’ Instructions emphasized accuracy since research suggests that latency varies less with differences in accuracy when instructions emphasize speed (Ratcliff and Rouder, 1998). Instructions encouraged subjects to respond quickly after reaching a decision in the hope of reducing variance due to

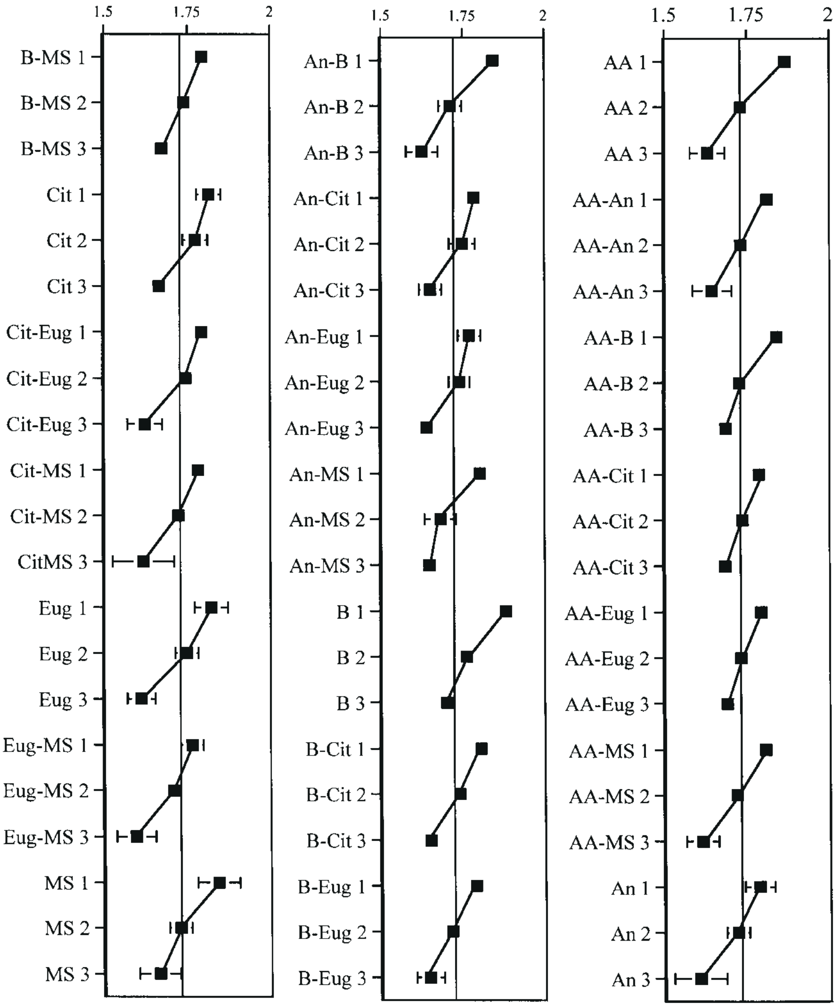

Figure 1 Log ratings of perceived intensity (± 1 SEM) for the 21 odorants, six single and 15 mixtures, at three concentrations (1 = highest, 2 = 5-fold dilution of 1, 3 = 5-fold dilution of 2). Two odorants separated by a dash denote a binary mixture, e.g. AA-An represents the binary mixture of amyl acetate and anethole. The middle point of each triad represents the rating for the concentration used for discrimination, and the vertical lines represent the rating for the standard.

factors such as lapses in attention. Instructions also encouraged subjects to use caution before responding 'same'. The results section will explain the reason for this.

At least 1 min separated successive trials. During this interval, the experimenter provided verbal feedback ('correct' or 'incorrect'), evacuated the tubes and placed the bottles for the next trial. Bottles were shaken thoroughly before placement to facilitate saturation of the head-space. The computer determined the order of trials by selecting randomly, without replacement, from a list of all trials in the block.

Blocks consisted of 57 trials, i.e. 57 odor-pairs, as follows: (i) 15 trials, all pairwise combinations between the different single odorants, e.g. AA versus An; (ii) 30 trials, all combinations of single odorants with binary mixtures of the odorant in question, e.g. AA versus AA mixed with An; and (iii) 12 trials, two trials of each odorant paired with itself, e.g. AA versus AA. Blocks included two trials of each odor paired with itself for a reasonable ratio of nominally same to nominally different pairs. In other words, the methodology largely matched a roving-stimulus, same–different design as described by Macmillan and Creelman (1991), except that subjects received no nominally same pairs of mixtures (Macmillan and Creelman, 1991). The order in which subjects received the members of each odor-pair varied randomly across blocks. Each block required ~70 min to complete. If subjects completed more than one block in a single day, 45 min breaks separated the blocks. Each subject completed a total of 20 blocks, or 20 judgements for each of the unmixed and mixed *different* pairs, and 40 judgements for each of the *same* pairs.

General notes on data treatment

This paper will focus on differences in latency rather than absolute latency. To demonstrate reliability of differences, 95% confidence intervals appear in most tables. Since 95% confidence intervals assume a normal distribution, and since distributions of latency tend to be skewed positively (Woodworth and Schlosberg, 1954), the geometric mean serves as the measure of central tendency. This consideration also prompted a log-transformation of distributions of latency before further analysis (see Figure 3).

By the standards of work in other sensory modalities, where collection of data proceeds more quickly, subjects contributed relatively few trials per odor-pair. Under these conditions, data from individual subjects for individual pairs achieved limited stability, but subjects could evaluate a relatively large number of pairs. Since this study sought primarily to examine some basic issues concerning measurement rather than provide data of archival quality on any single pair, a large stimulus-sample seemed appropriate. Given the above considerations, analyses are based on data averaged across subjects and/or odor-pairs.

Results

Measurement of absolute threshold

Log dilution to threshold from stock solutions used to study discrimination (mean of subject means ± SEM) follow: 4.16 ± 0.15, 4.15 ± 0.21, 3.78 ± 0.22, 3.68 ± 0.16, 5.43 ± 0.17 and 3.97 ± 0.24 for An, B, Cit, Eug and MS, respectively. Thus, the stimuli lay well and, with the exception of eugenol, more or less uniformly above threshold.

Ratings of perceived intensity

Each triad of points connected by line-segments in Figure 1 represents average ratings of perceived intensity for an odorant at three concentrations. The middle point represents the rating for the concentration used in discrimination and the vertical lines represent the rating for the standard. The concentrations used for discrimination proved well matched in perceived intensity for the current subjects.

Discrimination

Does latency reflect accuracy? In order for latency to discriminate to serve as a measure of similarity, different stimuli must elicit different latencies. Do difficult judgements require more time than less difficult judgements? Figure 2 shows that latency decreased as accuracy increased. Proportion correct for the 51 conditions (45 nominally different pairs and six nominally same pairs) appear in bins of 0.1, with average latency plotted in the middle of bins. Latency decreased monotonically between the intervals of 50–60% correct and 90–100%. Latencies for correct responses and for all responses combined decreased by similar amounts (450 and 420 ms, respectively) over the range of accuracy.

A measure of the sensitivity of latency to differences in quality. Latency of responses could potentially give rise to an interval scale (Stevens, 1950), but differences in latency could give rise to a ratio scale. One can, for instance, compare nominally different pairs with nominally same pairs. For this subtraction to yield a scale, same and different pairs should give rise to different distributions of latency. This consideration motivated the introduction of a response-bias in favor of answering ‘different’, since bias can separate same and different distributions (Ratcliff and Hacker, 1981). The TSD bias statistic C_{sd} [see equation 6.5 from (Macmillan and Creelman, 1991)] is positive if subjects display a bias toward responding ‘same’ and negative if subjects display a bias toward ‘different’. On average, subjects emitted 81% correct responses to different pairs, but only 60% correct responses to same pairs. This yields a C_{sd} of –1.48, indicating that subjects displayed on average a bias toward responding ‘different’.

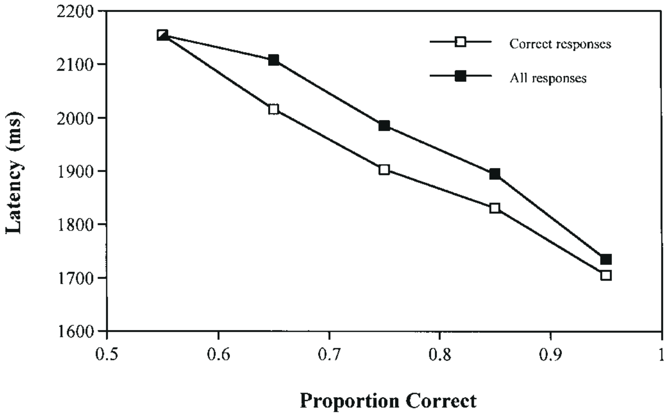

Figure 2 Latency as a function of proportion correct with (filled symbols) and without (open symbols) incorrect responses included. The latencies of responses were normalized (scaled by constants) so as to match mean latency for each subject to mean latency across subjects (normalization served to correct for individual differences in absolute latency that might otherwise obscure general trends). The data for each subject were then sorted by proportion correct, i.e. the 51 odor-pairs were arranged in ascending order according to accuracy. Average latency within bins of 0.1 was computed for each subject (subject 6 emitted slightly fewer than 50% correct responses on some same-pairs; in order to obviate the need for an additional bin, these data were entered into her 0.5–0.6 bin). The resulting latencies were averaged across subjects and plotted in the middle of each bin.

Figure 3 shows that all subjects responded more slowly to pairs of the same odors than to pairs of unmixed, different odors. Latency *d*′s expressed the separation between the means of the same and different distributions in units of their standard deviations. In other words, *d*′ represented the difference between a ‘comparison’ distribution, where the odor-pairs matched in quality, and another distribution, where the odor-pairs differed in quality. For example, the distribution of responses to AA versus AA and An versus An served as the comparison for the distribution of responses to AA versus An.

Latency-based *d*′s were computed from ROC curves. These plots show the proportion of latencies to evaluate

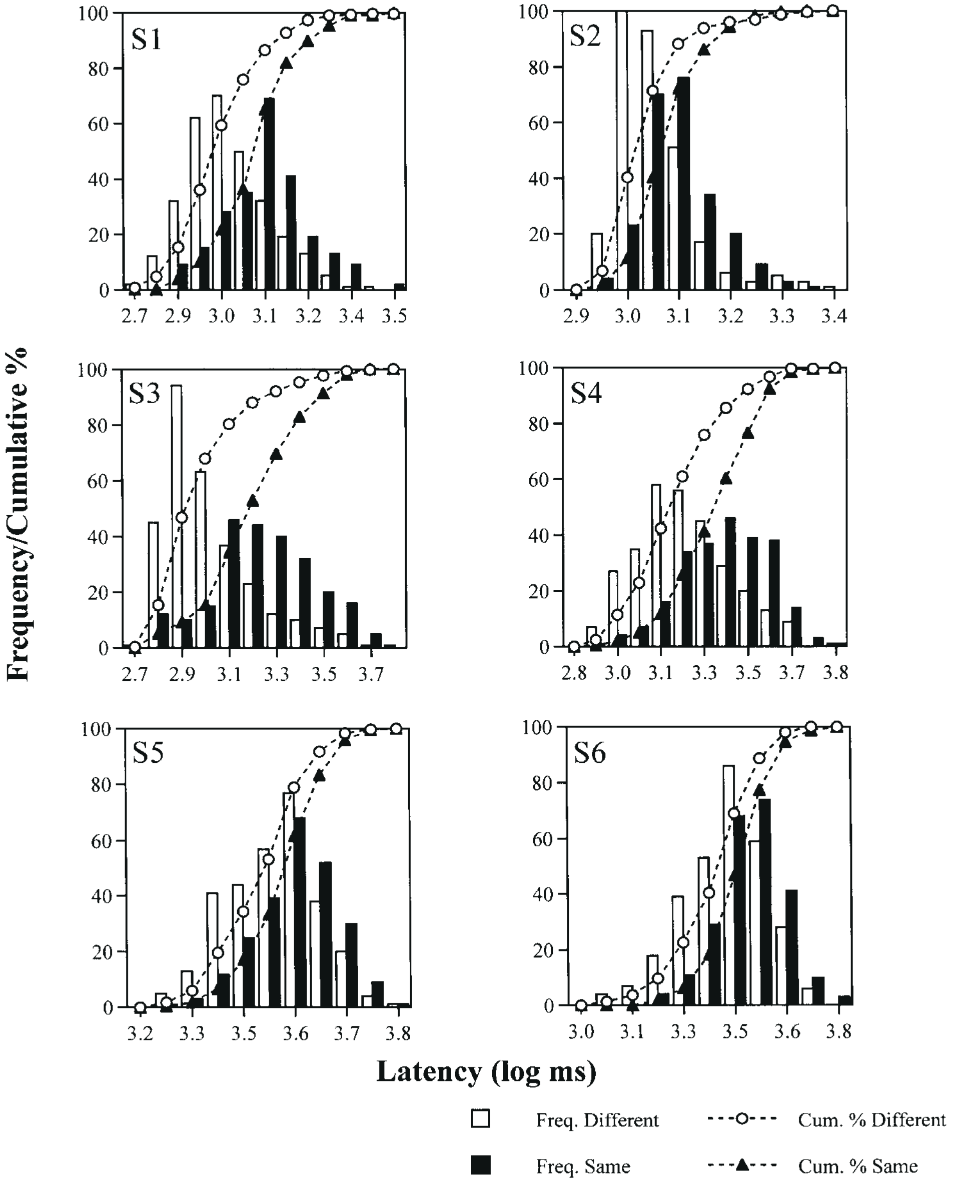

Figure 3 Latency to make same–different judgements (log ms) for all pairs of the same odorants (filled histograms) and for all pairs of different single odorants (open histograms). The curves represent the corresponding cumulative frequency distributions. Each panel represents responses of a single subject.

different pairs (y-axis) versus the proportion of latencies to evaluate same pairs (x-axis) that fall below a number of equally spaced criterion-latencies. Treating the data from Figure 3 in this way produced Figure 4. If the same and different distributions had coincided, the plots would have fallen along the main diagonals. The degree of curvature toward the upper-left corner represents the degree of separation between distributions. The area under an ROC curve, $p(A)$, equals the proportion correct in a 2-AFC task by an unbiased observer (Green, 1966; Geschieder, 1997). We used this finding to 'convert' $p(A)$s to d's using standard tables (Macmillan and Creelman, 1991).

Latency for correct responses and errors for same and different pairs. In general, judgements of 'same' required more time than judgements of 'different' for both nominally same and different pairs (Table 1). Nevertheless, Figure 5 shows that including the incorrect responses had little effect on the distributions for pairs of the same odor and for pairs of unmixed, different odors. This also held true for unmixed chemicals paired with mixtures. Since including errors apparently had no major effects, and since excluding incorrect responses to some nominally different pairs would have left fewer judgements with which to compute ROC curves, analyses included incorrect responses.

A measure of sensitivity based on errors of discrimination. MacMillan and Creelman provide tables of d' given hit and

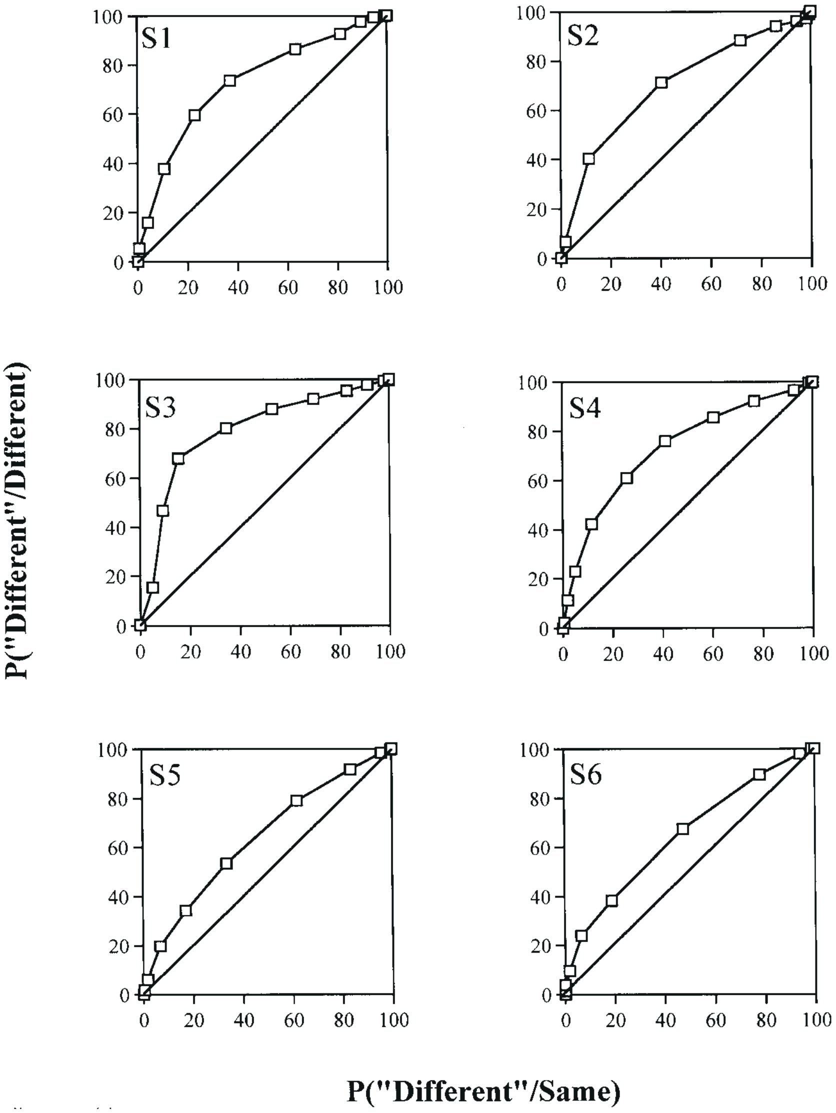

Figure 4 ROC curves based on latencies in Figure 3. The *x*-axis represents the percent of the distribution for pairs of the same odorants quicker than various criterion latencies, or *p*('Different'/Same) in the terminology of signal detection theory. The *y*-axis represents the percent of the distribution for pairs of different odorants quicker than various criterion latencies, or *p*('Different'/Different). If the distribution of latencies to evaluate same pairs did not differ from the distribution of latencies to evaluate different pairs, then the curves would have fallen on the main diagonals. The more the two distributions differed, the more the curves bowed out toward the upper-left corner. Each panel represents responses of a single subject.

false alarms rate in a same–different paradigm (MacMillan and Creelman, 1991). For now, analysis will assume that subjects used a differencing strategy. Data from the second study will support this choice. An example of how hits and false alarms were defined follows: the proportion of 'different' responses to AA versus An served as the hit rate, or *p*(Hit), for that pair, whereas the proportion of 'different' responses to AA versus AA and to An versus An served as the false alarm rate, or *p*(FA). To calculate *d*′, Probabilities of 1.0 were entered as 0.99.

d′*s between single odorants versus d*′*s between single odorants and mixtures.* Single odorant A should differ more in quality from single odorant B than either differs from the binary mixture A–B. Subjects should demonstrate higher *d*′s (shorter latencies) for discriminations of A versus B than for discriminations of A versus A–B or B versus A–B. In each pair of points connected by a line-segment (Figure 6), the upper point represents a *d*′ for a discrimination between two single odorants, and the lower point represents a *d*′ for a discrimination between a single odorant and a mixture. For all but one of the pairs of *d*′s based on latency, unmixed versus unmixed yielded a higher *d*′ than unmixed versus mixture. The same general pattern held for the *d*′s based on errors of discrimination, Pearson's r (43) = 0.88, $P < 0.001$. Both latency and error-based *d*′s therefore reflect differences between single odorants and mixtures.

Table 1 Latency (95% confidence interval of the geometric mean in s) and accuracy of responses

Subject	Condition		
	Sames[a]	Mixed versus unmixed	Unmixed versus unmixed
S1			
All	1.18–1.27	1.06–1.12	0.93–1.00
Correct	1.21–1.33	1.02–1.07	0.90–0.97
Incorrect	1.10–1.24	1.23–1.36	1.28–1.56
% correct	62.5	79.7	92.3
S2			
All	1.15–1.19	1.11–1.14	1.05–1.09
Correct	1.15–1.20	1.11–1.14	1.04–1.08
Incorrect	1.11–1.19	1.19–1.28	1.00–1.30
% correct	75.4	81.0	96.3
S3			
All	1.45–1.63	1.25–1.35	0.92–1.10
Correct	1.52–1.73	1.13–1.23	0.89–0.97
Incorrect	1.27–1.56	1.58–1.79	1.51–2.46
% correct	62.5	73.0	94.3
S4			
All	2.07–2.29	1.73–1.85	1.45–1.60
Correct	2.00–2.08	1.61–1.73	1.36–1.50
Incorrect	2.07–2.42	2.14–2.40	2.08–2.65
% correct	57.1	75.0	87.7
S5			
All	3.46–3.67	3.30–3.44	3.05–3.24
Correct	3.30–3.55	3.29–3.45	2.98–3.18
Incorrect	3.60–3.95	3.37–3.61	3.35–3.72
% correct	57.5	71.8	84.7
S6			
All	2.81–3.00	2.62–2.74	2.38–2.55
Correct	2.82–3.10	2.48–2.63	2.28–2.46
Incorrect	2.74–3.00	2.91–3.11	2.77–3.13
% correct	45.0	70.5	81.3

[a] Total number of responses per subject as follows: sames, 240; mixed versus unmixed, 600; unmixed versus unmixed, 300.

Resolution among unmixed different pairs. Discriminative performance approaches an asymptote once odors differ beyond a certain extent, and error-based measures lose the ability to resolve differences among odor-pairs. Can latency provide better resolution? The average difference among the 15 *d*′s between unmixed odor-pairs was computed in two ways. First, resolution was computed in terms of between-subjects variance, i.e. *d*′ for AA versus An minus *d*′ for AA versus B divided by the average between-subjects standard error for the two pairs. Latency-based *d*′s differed by an average of 1.77 ± (SEM) 0.15 standard errors, whereas error-based *d*′s differed by an average of only 1.15 ± 0.09 standard errors. Second, resolution was computed in terms of percent differences. Latency-based *d*′s differed by 29.63 ± 2.60% on average. The highest exceeded the lowest by a factor of 3.32. Error-based *d*′s differed by 16.43 ± 1.04% on average. The highest exceeded the lowest by a factor of 1.66. With the same number of responses from the same number of subjects, the index of similarity based on latency provided better resolution between pairs of unmixed odors.

Do mixtures fall on a straight line connecting their components? If the qualities of binary mixtures fall between those of their components in 'odor-space', then the data should display additivity: adding the discriminability of A versus the mixture A–B to that of B versus A–B should return a value equal to A versus B. A value greater than A versus B means more than one dimension will be required to locate mixtures with respect to their components.

The bars in Figure 7a represent the average (across odorants) percent deviation from additivity for latency-based *d*′s, computed as follows:

$$\text{percent deviation} = 100[d'_{\text{components}} - (d'_{\text{mix1}} + d'_{\text{mix2}})]/d'_{\text{components}}$$

where $d'_{\text{components}}$ represents the *d*′ for the unmixed components versus one another, and d'_{mix1} and d'_{mix2} represent the *d*′s for a binary mixture versus its first and second components, respectively. Perfect additivity produces a 0% deviation. Since the 95% confidence intervals (error bars) for all subjects bracket zero, additivity held (additivity held across subjects as well). To the extent that latency to discriminate reflects similarity of quality, this finding suggests binary mixtures fall on straight lines connecting their components.

The bars in Figure 7b represent percent deviation from additivity for error-based *d*′s. All subjects demonstrated negative deviations, and the 95% confidence-intervals for five out of six subjects failed to bracket zero. The sum of the discriminabilities of the mixtures versus their components exceeded the discriminability of the components versus one another. To the extent that *d*′s based on errors reflect similarity of quality, this finding suggests binary mixtures may fall between their components on one dimension, but differ from their components on at least one other dimension.

Individual differences. The 15 pairwise correlations between the six subjects, across all 45 latency-based *d*′s, were positive, though three failed to reach statistical significance ($P < 0.05$, Table 2a). All correlations were positive for *d*′s based on errors of discrimination as well (Table 2b), though four failed to reach significance. Both dependent variables showed that one subject's discriminative 'odor-space' resembled those of others to some extent. However, since the responses of one subject predicted just 16% of the variance in the responses of another on average, substantial individual differences in 'odor-spaces' may also exist. Table 1 shows that substantial individual differences also exist in absolute latency and accuracy.

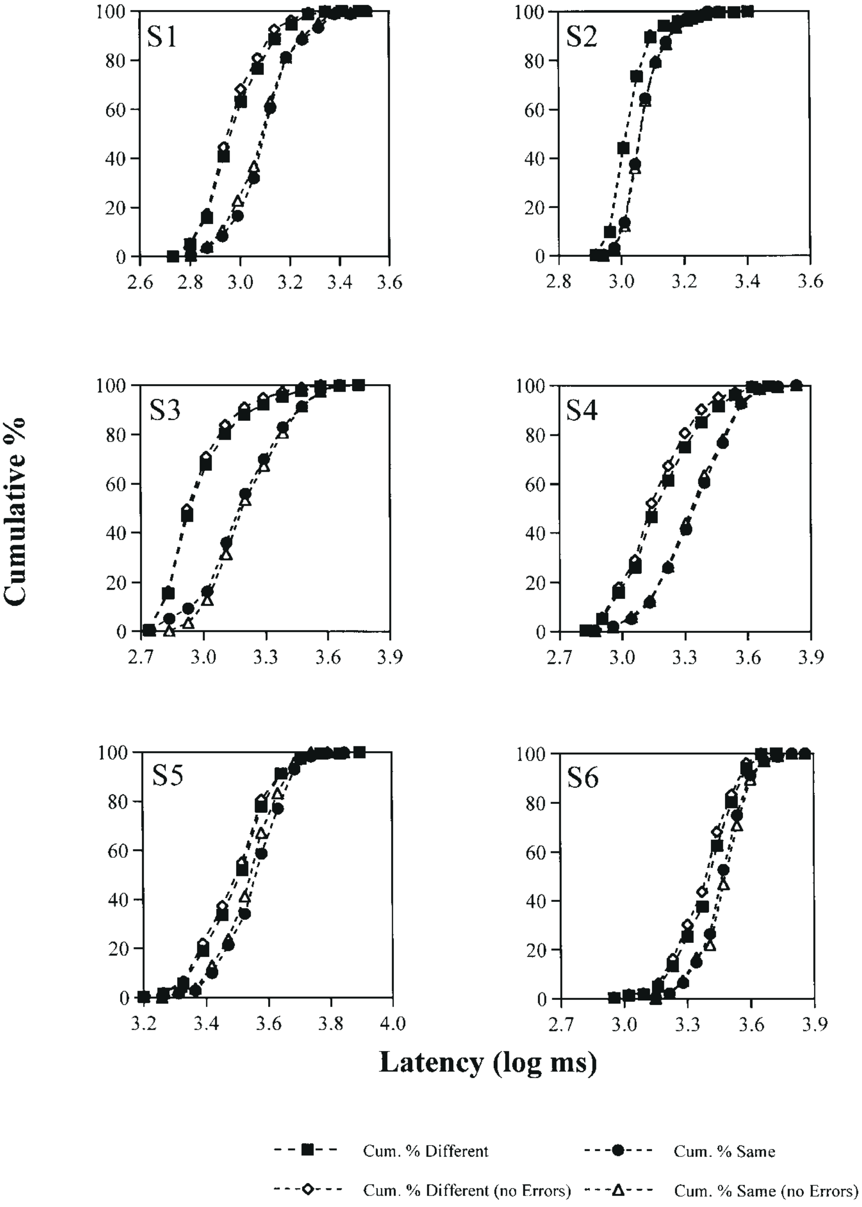

Figure 5 Cumulative frequency distributions for all pairs of different and all pairs of the same odorants. Filled symbols (squares for different and circles for same) represent distributions that include incorrect responses. Open symbols (diamonds for different and triangles for same) represent distributions that exclude incorrect responses. Each panel represents responses of a single subject.

Study 2

Can measured discriminabilities predict discriminabilities not yet measured? Multidimensional scaling (MDS) of latency-based *d*′s among the six unmixed chemicals from study 1 led to an 'odor-space'. Binary mixtures were 'placed' in the space through linear interpolation between unmixed chemicals. The distances between mixtures in the odor-space served as predicted relative discriminabilities among binary mixtures.

Since this study included a smaller number of pairs, subjects emitted more trials per condition. This allowed the calculation of latency-based *d*′s both with and without errors included. Latency *d*′s might be somewhat higher after exclusion of the errors, but will the *relative* values change? If so, then experimenters should collect enough trials for analysis without the incorrect responses. If not, then it may not matter as much how experimenters treat errors, as long as they treat errors consistently.

To our knowledge, no one has empirically determined the proper TSD model for same–different discriminations of odor-quality. This study will fill this gap by plotting full ROC curves based on errors of discrimination. One can do this without experimentally manipulating subjects' criteria but by collecting ratings of confidence (Macmillan and

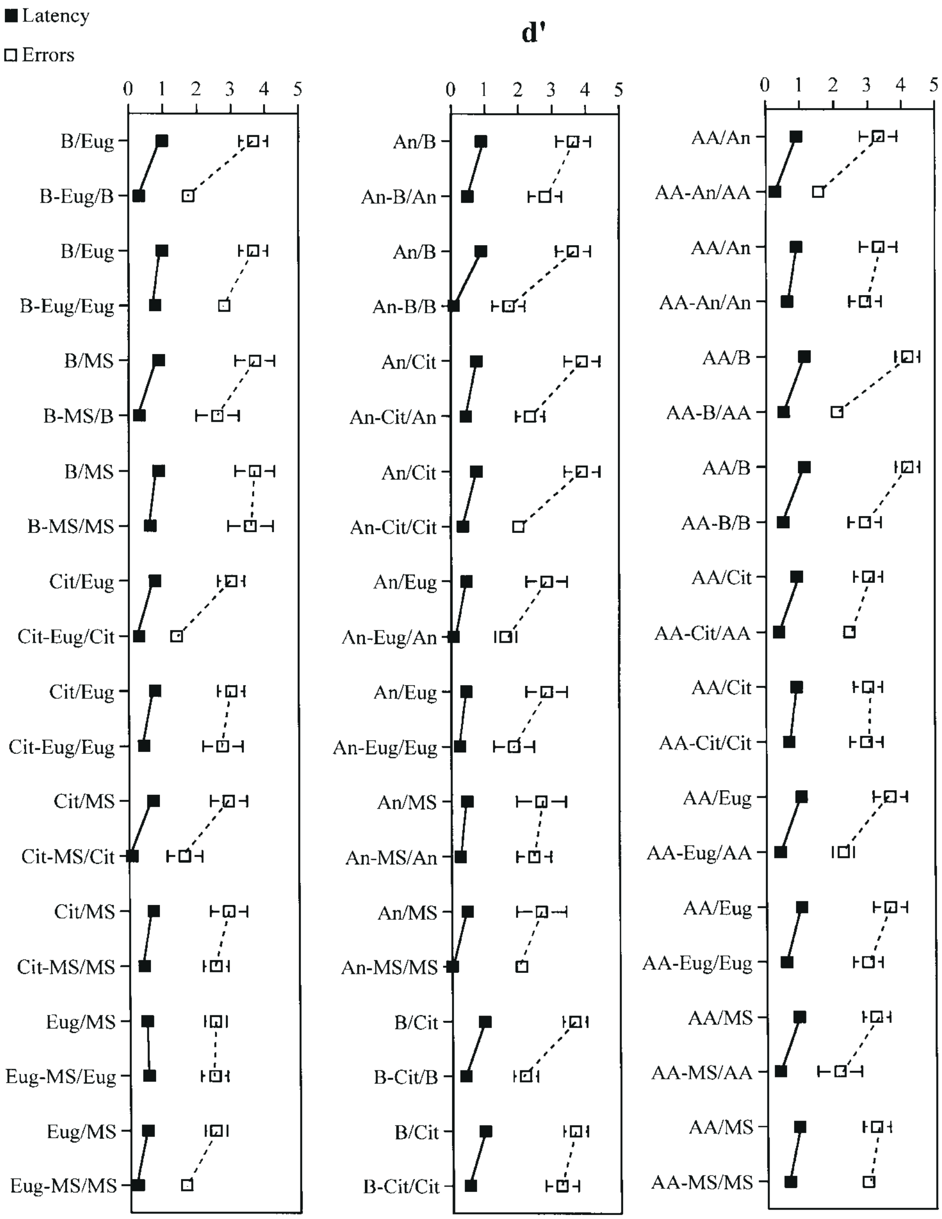

Figure 6 Mean values (across subjects, ±1 SEM) of *d*'s based on latency (filled symbols) and *d*'s based on errors of discrimination (open symbols). A slash mark means 'versus'. Thus, AA/An represents a *d*′ for amyl acetate versus anethole. A hyphen denotes a binary mixture. Thus, AA-An/AA represents a *d*′ for the binary mixture of amyl acetate and anethole versus amyl acetate.

Creelman, 1991). Different levels of confidence can represent different 'criteria'. A strategy based on independent observations leads to ROC curves symmetric with respect to the negative diagonal, whereas a strategy based on differencing leads to ROC curves asymmetric with respect to the negative diagonal (Macmillan and Creelman, 1991; Irwin and Francis, 1995; Francis and Irwin, 1997).

Methods

Subjects

Four young, non-smoking females (ages 19–31 years) participated in ~11 h of testing each. Two were laboratory assistants, two were paid volunteers. All had served in study 1 (subjects 1–4, with the same designations used in both studies).

Stimuli

The stimuli included three binary mixtures, prepared as in study 1: amyl acetate mixed with benzaldehyde (AA–B), anethole mixed with citral (An–Cit), and eugenol mixed with methyl salicylate (Eug–MS).

Calculation of predicted relative discriminabilities

The 15 pairwise discriminabilities among the six unmixed odors from study 1, averaged across the four subjects who

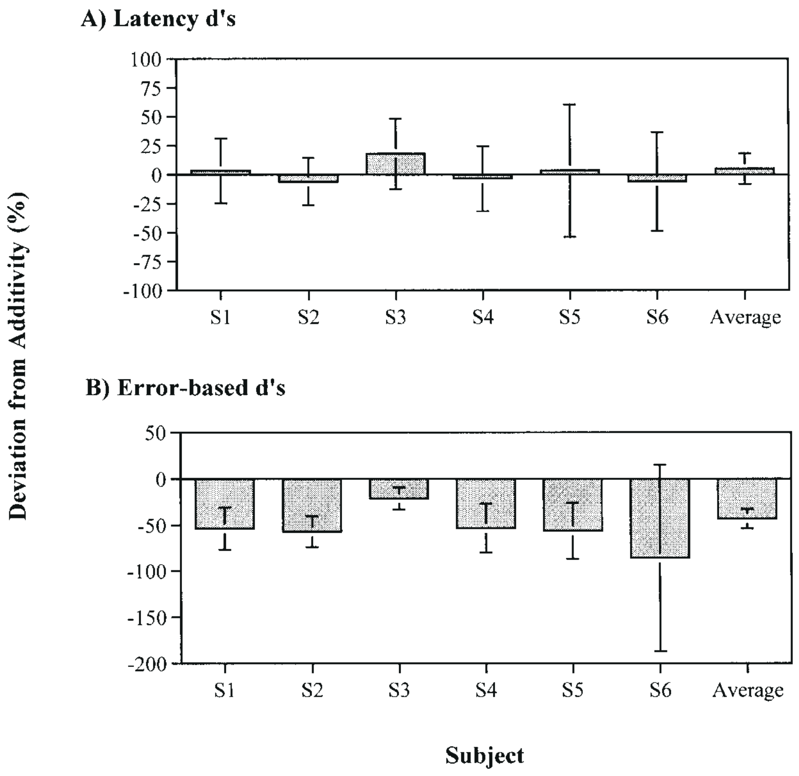

Figure 7 Average (across pairs, ±2 SEM) indices of additivity by subject and averaged across subjects for **(A)** latency-based *d'*s and **(B)** error-based *d'*s.

Table 2 Correlations (Pearson's *r*) between subjects for *d'*s

	S1	S2	S3	S4	S5	S6
Based on latency						
S1	1	0.52	0.58	0.43	0.45	0.38
S2		1	0.51	0.34	**0.24**	0.31
S3			1	0.55	0.58	0.31
S4				1	0.61	**0.18**
S5					1	**0.14**
S6						1
Based on errors						
S1	1	0.42	0.50	0.37	0.36	0.34
S2		1	0.32	**0.11**	0.37	**0.03**
S3			1	0.62	0.49	0.48
S4				1	0.55	**0.27**
S5					1	**0.12**
S6						1

Values in bold fail to reach significance (i.e. $P > 0.05$).

served in study 2 (because *d'*s from study 1 were based on relatively few trials per odor-pair, they did not allow reliable predictions for individual subjects), were submitted to two-dimensional Guttman MDS analysis. Solutions accounted for 99 and 89% of the variance for the latency-based and error-based *d'*s, respectively. The three mixtures were 'placed' in the spaces by linear interpolation, i.e. mixtures were located halfway along line-segments connecting their components. If the 'odor-space' is valid for the current sample of subjects, and if binary mixtures of equally intense odors fall about halfway between their components, the distances between the mixtures should predict their relative discriminabilities.

Procedure

The procedure followed that of study 1, except for the following: (i) subjects received no feedback in study 2. Since the stimulus set for study 2 included fewer odors (three instead of 21), feedback might have been more likely to help subjects focus on non-qualitative cues such as small differences in intensity. (ii) After responding 'same' or 'different', subjects rated confidence in responses by pressing number-keys on a standard keyboard. Subjects rated confidence on a scale from one to five: '1' reflected very low confidence, '3' reflected moderate confidence, and '5' reflected very high confidence.

Blocks consisted of 54 trials: (i) 36 all pairwise combinations of the mixtures, 12 repetitions each, and (ii) 18 mixtures paired with themselves, six repetitions each. The order in which subjects received the members of an odor-pair was counterbalanced within blocks. A subject completed nine blocks (one for practice, eight for analysis), for a total of 96 and 48 judgements for each *different* and *same* pair, respectively.

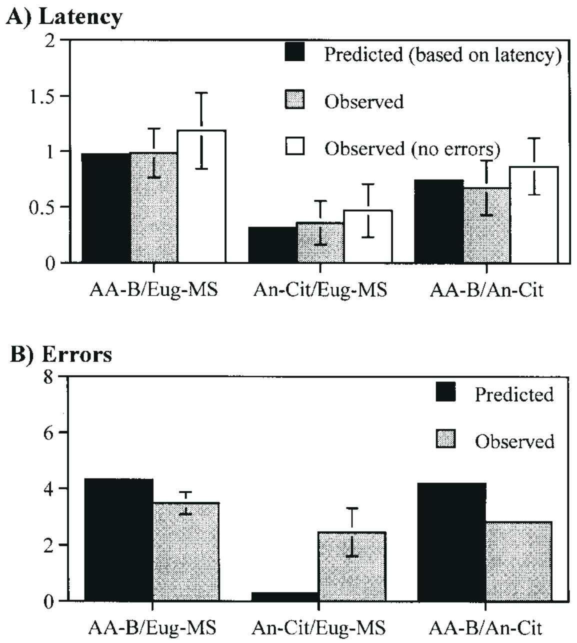

Figure 8 **(A)** Predicted relative discriminabilities (black bars) and observed latency-based *d*'s (gray bars with incorrect responses included, white bars with incorrect responses excluded) among three binary mixtures (mean across subjects, ±1 SEM). Predicted values were scaled such that their mean equaled the mean of the observed *d*'s with the errors included. **(B)** Predicted relative discriminabilities (black bars) and observed error-based *d*'s (gray bars) among three binary mixtures (mean across subjects, ±1 SEM). Predicted values were both scaled such that their mean equaled the mean of the observed error-based *d*'s.

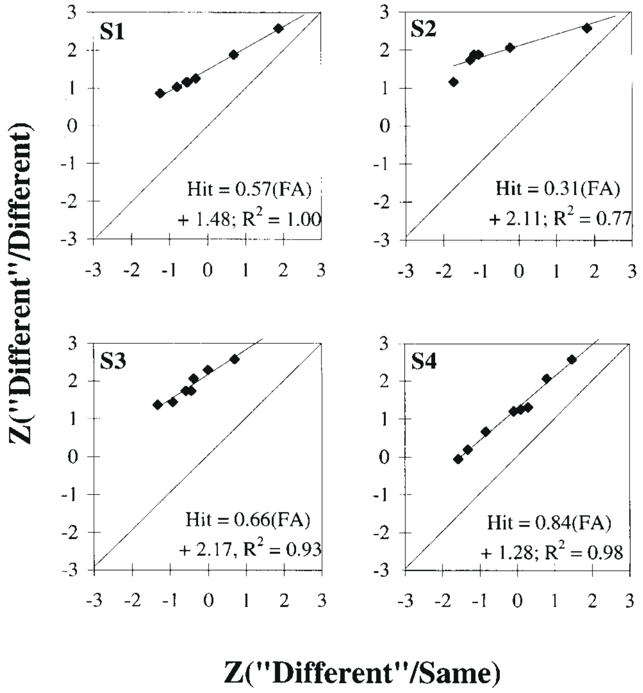

Figure 9 ROC curves for the single odor-pair AA–B versus Eug–MS (shown for illustration), based on ratings of confidence, in normal deviate coordinates. The *x*-axis depicts *z*(FA) and the *y*-axis depicts *z*(Hit) at each level of confidence (see text for more details). Hit/false-alarm pairs that include probabilities of 0.0 or 1.0 do not appear in the graphs, since normal deviate values of 0 and 1 are undefined. The graphs contain <10 points for this reason. Note that the curves are approximately linear in these coordinates with slopes of <1. Each panel represents data for a single subject.

Results

Observed versus predicted discriminabilities

Figure 8a illustrates predicted relative discriminabilities versus observed discriminabilities based on latency. Observed values agreed quite well with relative predicted values. The *d*'s computed with the errors excluded showed the same pattern, but nominally exceeded the *d*'s computed with the errors included. However, the difference was not statistically reliable. Figure 8b illustrates predicted relative discriminabilities versus observed error-based discriminabilities. Observed values did not agree well with predicted values.

Fits of independent-observations and differencing models to ROC curves

The five-point ratings of confidence for sames and differents combined to give a 10-point scale, with 1 = very high confidence 'same' and 10 = very high 'different'. A program plotted the cumulative proportion of nominally same pairs versus cumulative proportion of nominally different pairs at each confidence-level. Twelve ROC curves, one for each odor-pair for each subject, were plotted. A curve represented 96 trials of given different pair and 48 trials of each associated same pair, or a total of 188 trials.

In normal deviate coordinates, same–different ROC curves plot as roughly straight lines (Figure 9). An asymmetric ROC curve has a slope of <1, whereas a symmetric ROC curve has a slope of ~1 (Macmillan and Creelman, 1991). Table 3 provides a summary of linear fits for all pairs and all subjects. The average slope equaled 0.63, with a 95% confidence interval from 0.50 to 0.75, clearly less than 1.

Linear fits described the ROC curves well, but they minimize deviations along the *y*-axis only. Since both hit- and false-alarm rates qualify as dependent variables, the curves were also fit through maximum likelihood estimation (MLE). Software developed by John Irwin and co-authors was used to fit both differencing and independent-observations models. Since the software does not accommodate 10 intervals of confidence, the data was binned into five intervals such that ratings of 1 and 2 comprised the first interval, ratings of 3 and 4 comprised the second interval, and so on.

Results of fits appear in Table 4. The chi-squared values indicate the quality of the fits: the lower the value, the better the fit. In all but one case, the differencing model gave a lower chi-squared than the independent-observations model. No curve showed a statistically significant deviation from the differencing model, and the summed chi-square did not approach significance. Significant deviations from the independent-observations model occurred for seven out of 12 ROC curves, and the summed chi-square reached significance at $P < 0.001$. This result, like the slopes of the curves in normal deviate coordinates, shows that the ROC curves are not symmetric. A differencing model fit best.

The MLE software also calculated estimates of standard errors of d's for individual subjects and odor-pairs (Table 4). Standard errors of d' were on average (across subjects and odor-pairs) only 7.60% of the values of d'. Thus, 188 trials afforded considerable precision in the estimates of similarity between odorants.

Table 3 Slope of linear fits of ROC curves in normal deviate coordinates

Subject	Pair	Slope	r^2
1	1[a]	0.65	0.97
	2	0.57	1.00
	3	0.70	1.00
2	1	0.22	0.78
	2	0.31	0.77
	3	0.53	0.99
3	1	0.52	0.92
	2	0.66	0.93
	3	0.85	0.84
4	1	0.97	0.98
	2	0.84	0.98
	3	0.72	0.99
Average		0.63	0.93

[a] 1 = An–Cit/Eug–MS, 2 = AA–B/Eug–MS, 3 = An–Cit/Eug–MS.

Discussion

Do discriminations between similar odor-pairs require more time than discriminations between less similar odor-pairs?

Study 1 showed that latency decreased monotonically with accuracy (Figure 2). The relationship between accuracy and latency became manifest again in the strong correlation between d's based on accuracy and those based on latency. Most importantly, a manipulation of quality led to clear changes both in accuracy and latency: subjects required more time and made more errors (demonstrated lower latency- and error-based d's) in discriminations between binary mixtures and their unmixed components than in discriminations between unmixed components (Figure 6). The results suggest that discriminability plays a role in determining response times in olfactory discriminations, as it does in the auditory and visual modalities (Woodworth and Schlosberg, 1954; Welford, 1960; Vickers, 1980; Luce, 1986; Sinnott, 1989; Sinnott *et al.*, 1997). In short, since latency conveys information regarding discriminability, it meets an essential requirement for a potential measure of differences in odor quality.

The relative values of error- and latency-based d's showed reasonable agreement, but the absolute values of error-based d's were higher. Error-based d's estimate differences between theoretical distributions of sensation, whereas latency-based d's represent differences between empirical

Table 4 Summary MLE fits to confidence ROCs

Subject	Pair	$p(A)$	Differencing strategy					Independent-observations strategy				
			d'	SE	chi-sq.	df	P	d'	SE	chi-sq.	df	P
1	1[a]	0.72	1.95	0.14	6.62	3.00	0.09	1.52	0.17	13.43	3.00	0.00
	2	0.90	3.28	0.13	3.13	3.00	0.37	2.54	0.16	13.07	3.00	0.00
	3	0.81	2.54	0.13	1.29	3.00	0.73	2.00	0.15	6.96	3.00	0.07
2	1	0.99	5.10	0.14	6.22	3.00	0.10	3.89	0.19	10.13	3.00	0.02
	2	0.97	4.31	0.14	4.37	3.00	0.22	3.31	0.17	7.60	3.00	0.06
	3	0.86	2.85	0.13	5.27	3.00	0.15	2.27	0.16	9.17	3.00	0.03
3	1	0.83	2.63	0.13	3.00	3.00	0.39	2.05	0.15	13.50	3.00	0.00
	2	0.95	4.02	0.19	2.69	3.00	0.44	2.98	0.16	11.18	3.00	0.02
	3	0.90	3.25	0.14	5.06	3.00	0.17	2.53	0.16	9.74	3.00	0.02
4	1	0.53	0.58	0.23	1.25	3.00	0.74	0.53	0.30	1.02	3.00	0.80
	2	0.80	2.47	0.14	1.30	3.00	0.73	1.91	0.15	7.08	3.00	0.07
	3	0.84	2.69	0.14	1.12	3.00	0.77	2.09	0.15	6.08	3.00	0.11
Sum					41.32	36	>0.25			108.96	36	<<0.01

[a] 1 = An–Cit/Eug–MS, 2 = AA–B/Eug–MS, 3 = An–Cit/Eug–MS.

distributions of latency. Thus, absolute values would not agree, except by chance.

A measure of similarity based on latency

Various investigators have constructed operator characteristic curves based on latency (Bindura *et al.*, 1965; Carterette *et al.*, 1965; Norman and Wickelgren, 1969; Emmerich *et al.*, 1972; Luce, 1986). For example, Katz assigned the fastest 'same' and 'different' responses scores on opposite ends of a scale, and plotted operating characteristics according to a method analogous to that used in the current study to plot confidence ROCs (Katz, 1970). Such measures cannot resolve differences beyond the point of asymptotic performance.

The current study employed a measure based on the difference between the distributions of latency to evaluate nominally same and nominally different pairs, i.e. classified responses according to the physical stimulus alone. This measure can potentially reflect differences in quality beyond the level of asymptotic discriminability and can potentially provide ratio-level measurement. However, the approach will work only if nominally same and different pairs give rise to different distributions of latency.

The role of bias

The majority of past studies of same–different discriminations, mostly between strings of letters or digits, have found that same pairs typically take less time to evaluate than different pairs (Luce, 1986). Some have proposed that a separate mechanism, a fast 'identity matcher', processes same pairs (Bamber, 1969; Taylor, 1976), or that subjects encode same pairs differently (Proctor, 1981). However, Ratcliff and Hacker have shown that subjects told to respond 'same' only when very certain emitted 'same' responses more slowly than most 'different' responses, whereas subjects told to respond 'different' only when certain emitted 'different' responses more slowly (Ratcliff and Hacker, 1981).

Various authors have proposed that response time in a detection or discrimination task decreases with increasing distance between sensation and criterion along the decision axis (Bindura *et al.*, 1965; Norman and Wickelgren, 1969; Emmerich *et al.*, 1972; Espinoza-Varas and Watson, 1994), although subjects probably do not compute this distance consciously. In the current study, a bias in favor of responding 'different' was introduced through instruction and a higher *a priori* probability of the occurrence of nominally different pairs. Under these conditions, subjects' same–different criteria should lie closer to the distribution of differences in sensation due to nominally same pairs. In the context of the hypothesis described above, it makes sense that subjects took longer to evaluate same pairs.

The preceding discussion suggests that investigators can use bias to separate distributions of latency in same–different discriminations of odor quality. Systematic manipulations of bias (Ratcliff and Hacker, 1981) could prove this conclusively. If so, one could always calculate latency-based *d*′s as defined here, but the exact values would depend on the degree of bias. From the standpoint of measurement, bias poses no problems (and, as the preceding discussion suggests, could prove useful) as long as the degree of bias is known and remains constant. One could adjust subjects' criteria to a desired (constant) value *via* instruction and feedback. Specifically, one could coach subjects to a desired ratio of false 'sames' to false 'differents'.

Speed–accuracy trade-off

Latency does not change as much with accuracy when instructions emphasize speed (Ratcliff and Rouder, 1998). Since the slope of the latency versus accuracy function can change with emphasis on speed versus accuracy, latency-based *d*′s could also change. Some data from the current study may illustrate this point. Subjects 1, 2 and 3 took longer to evaluate same pairs in the second study, but achieved greater accuracy (compare Tables 1 and 5). Since they also demonstrated high accuracy in evaluating different pairs, particularly subjects 2 and 3, they probably did not achieve greater accuracy with the sames at the expense of the differents, but achieved greater accuracy by taking more time.

In both studies, subjects were asked not to sacrifice accuracy for speed. Nevertheless, some subjects may have rushed a bit in study 1. Subjects were told to respond quickly after they had reached a decision in order to avoid outlying responses due to lapses in attention. However, it seems that mention of response times may have influenced subjects. In the future, perhaps instructions should encourage subjects to give their best answer (given their current degree of bias toward responding 'different') without mention of response time.

From the standpoint of measurement, olfactory scientists can deal with bias and speed–accuracy trade-off. In principle, investigators can emphasize only accuracy and set bias to a fairly constant value through instructions and feedback. The process will probably take some effort, but a rich literature on response times in other modalities can serve as a guide (Luce, 1986; Smith and Vickers, 1988; Ratcliff and Rouder, 1998).

Resolution among pairs of unmixed odorants by latency- and error-based *d*′s

The first study showed that, with the same number of subjects and judgements, *d*′s based on latency provided better resolution among pairs of unmixed odors than *d*′s based on accuracy. Latency-based *d*′s provided superior resolution both in terms of between-subject variability and percent difference.

Although latency-based *d*′s provided better resolution of differences between *odor-pairs*, error-based *d*′s sometimes provided better resolution of differences between *individual*

Table 5 Latency (95% confidence interval of the geometric mean in s) and accuracy of responses in study 2

Subject	Condition			
	Sames[a]	AA–B/Eug–MS	AA–B/An–Cit	An–Cit/Eug–MS
S1				
All	1.82–1.96	1.44–1.60	1.73–1.91	1.77–1.94
Correct	1.83–1.98	1.38–1.53	1.62–1.80	1.64–1.88
Incorrect	1.69–2.00	1.83–2.26	2.00–2.33	1.96–2.08
% correct	75.0	87.5	72.9	62.5
S2				
All	1.32–1.38	1.17–1.25	1.27–1.35	1.17–1.23
Correct	1.31–1.37	1.17–1.25	1.24–1.33	1.17–1.23
Incorrect	1.34–1.55	1.21–1.41	1.31–1.46	1.24[b]
% correct	86.7	95.8	96.3	99.0
S3				
All	1.75–1.96	1.03–1.17	1.17–1.36	1.33–1.59
Correct	1.78–2.00	1.01–1.12	1.10–1.26	1.20–1.44
Incorrect	1.56–2.22	1.61–3.25	1.97–2.88	2.05–2.30
% correct	73.4	95.8	90.6	80.2
S4				
All	1.48–1.68	1.06–1.26	1.05–1.22	1.44–1.70
Correct	1.39–1.68	1.02–1.18	0.99–1.15	1.41–1.79
Incorrect	1.49–1.76	1.14–2.25	1.33–2.10	1.42–1.65
% correct	47.9	90.6	87.5	58.5

[a] Total number of responses per subject as follows: sames, 144; differents, 96.

[b] Single response.

odors. In study 1, some latency-based *d*′s between a mixture and an unmixed odorant did not differ from zero, whereas the corresponding error-based *d*′s did differ from zero. Latency-based *d*′s might serve best in the resolution of larger differences in quality, i.e. might extend the range of qualitative differences discrimination can resolve.

Do the qualities of binary mixtures fall between those of their components?

To a first approximation, additivity of discriminability held for latency-based *d*′s. One dimension could describe the relationship between mixtures and their components. Additivity did not hold for *d*′s based on errors of discrimination, as the sum of *d*′s between mixtures and their components exceeded *d*′s between unmixed components. One dimension failed to describe the relationship between mixtures and their components. Why do the two dependent variables provide different pictures?

Perhaps performance approached an asymptotic level in discriminations between unmixed chemicals, and the corresponding error-based *d*′s underestimated the differences among the odors of the unmixed stimuli. In other words, error-based *d*′s may have accurately represented the relative differences between mixtures and their components, but underestimated differences between unmixed components. If this explanation is correct, *d*′s based on latency may have displayed additivity due to superior resolution.

The unmixed odors in this study differed a great deal, but may not have produced asymptotic performance. More research would be needed to determine whether the above explanation is correct. Discriminations between unmixed odors and proportional mixtures, constructed by gradually diluting one odorant with another while keeping total perceived intensity constant (e.g. A versus a 90–10 mixture of A and B, A versus a 80–20 mixture of A and B, and so on) could shed light on this issue. Discriminability increases monotonically as one dilutes a chemical with another, at least over some range (Olsson and Cain, 2000). If performance approaches an asymptote earlier in the series than A versus unmixed B, then the difference between A and B probably falls beyond the asymptotic level. One could determine the limits of the resolution of latency and errors by comparing latency- and error-based *d*′s for a number of proportional series. If limited resolution does produce a failure of additivity, the index of additivity should equal 0 for all confusable pairs, and become increasingly negative for pairs beyond asymptotic discriminability.

Research on proportional mixtures might eventually lead to a quantitative model of the qualitative relationship between binary mixtures and their components (Olsson, 1994). The current study examined only 50–50 mixtures. A complete model would require measurements of a range of proportions (Olsson and Cain, 2000). Once olfactory science develops a model of binary mixtures, discrimination might help to model more complex mixtures, like those we experience in everyday life.

Observed versus predicted relative discriminabilities: a demonstration of internal consistency

Latency-based *d*′s among unmixed chemicals predicted the relative values of latency-based *d*′s among mixtures very well. Error-based *d*′s among unmixed chemicals did not predict the relative values of error-based *d*′s among mixtures. To some extent, the *d*′s based on errors may simply have had more noise. Further, as discussed previously, performance may have approached an asymptotic level for some pairs. *d*′s may have underestimated the differences between some chemicals, which could have led to a somewhat distorted 'odor-space' that produced poor predictions.

Predictive failure of error-based *d*′s does not detract from the success of the latency-based *d*′s. Their predictive ability is impressive for several reasons. First, predicted values were based on relatively few judgements (20 per subject for each nominally different pair). With more precision, prediction would surely improve. Second, the assumption that mixtures fall exactly halfway between their components was perhaps somewhat strong. Third, as a previous discussion suggests, subjects could have effected changes from study 1 to study 2 in criterion or emphasis on accuracy versus speed. Given these considerations, the ability of latency-based *d*′s to predict the relative discriminabilities of previously untested odor-pairs constitutes a strong demonstration of predictive validity.

As long as relative discriminabilities prove trustworthy, changes in absolute values pose few problems. If one wished to combine two data-sets, one could include some 'calibration-pairs' that both data-sets shared. The ratio of corresponding *d*′s from the two sets could serve to scale the values in one set. Differences in absolute values of latency-based *d*′s that might arise from slight differences in technique, criterion, speed–accuracy trade-off or other factors need not prevent accumulation of archival data if we can trust the relative values. So far, nothing suggests that we cannot.

Latency-based *d*′s with and without errors of discrimination included

Since the second study included fewer odor-pairs, subjects could emit more responses per pair. This allowed the construction of ROC curves both with and without errors included. Figure 8a shows that excluding errors resulted in nominally higher *d*′s. This makes sense, as subjects emitted incorrect 'different' responses more quickly than correct 'same' responses, and emitted incorrect 'same' responses more slowly than correct 'different' responses (Table 1). Excluding errors should separate the distributions and reduce their variance. However, the difference between the methods of analysis did not prove statistically reliable. Further, the two methods of analysis produced comparable relative values, i.e. the ratios of the *d*′s for the three odor-pairs stayed about the same whether or not one included errors. In short, it may not matter very much how an experimenter treats errors, as long as the method of analysis is consistent.

Models of same–different discriminations

Study 2 demonstrated that a differencing model fit the data better than an independent-observations model. In this study, subjects seemed to make relative rather than absolute judgements. In other words, subjects seemed to base their decisions regarding odor quality on 'simple' differences, as they do in discriminations of perceived orientation (Vogels and Orban, 1986), taste intensity (Irwin *et al.*, 1993) and loudness (Hautus *et al.*, 1994). This information matters a great deal for proper measurement, as incorrect application of an independent-observations model would have underestimated differences among pairs (Table 4).

Can subjects make absolute judgements regarding odor quality in other situations? Studies of odor identification suggest they can, since they can often name the exact substance that emitted an odor (Cain, 1979, 1982; Cain *et al.*, 1998). Further, that subjects can often find many olfactory 'notes' within an odor (Dravnieks, 1985) suggests that odor quality could in principle qualify as a multi-attribute sensation (although people may not naturally function in such an analytic mode). Complex, multidimensional stimuli can allow subjects to make absolute judgements (Irwin and Francis, 1995; Francis and Irwin, 1997).

One reason for subjects' failure to apply an independent-observations strategy may have lain in the roving-stimulus design of the current studies, i.e. a design where a number of odor-pairs are interleaved within a session (Macmillan and Creelman, 1991; Irwin and Francis, 1995). In this situation, the information needed to make absolute judgements may not be available (Macmillan and Creelman, 1991). The current studies did not employ a true roving design as described by Macmillan and Creelman since the pairs interleaved within a session did not differ by a constant amount along a given continuum. Nevertheless, future studies could determine if subjects can make absolute judgements in discriminations of odor quality in a fixed-stimulus design, i.e. a design where subjects evaluate a single pair of odors within a session.

Future studies could also determine if models assuming other decision-strategies could achieve even better fits. Recently, Dai and colleagues have shown that the independent-observations and differencing strategies assumed in the models used here may represent end-points along a continuum of strategies (Dai *et al.*, 1996). If so, then an examination of other, possibly intermediate, strategies could prove fruitful.

Individual differences in relative values

The correlations between subjects in Table 2 suggest that the discriminative odor-spaces of individuals resemble one another somewhat, but the weakness of the correlations suggests large individual differences may also occur. Given the small number of responses per pair, noise almost certainly contributed to the weakness of the correlations. However, data from study 2, which display more precision, also demonstrate individual differences (Table 4). Past studies of discrimination of odor quality have also found such differences (Jones and Elliot, 1975; Hummel *et al.*, 1992; Laska and Teubner, 1999).

Individual differences provide both a challenge and an opportunity. One must study a number of subjects to obtain a general picture of the similarity of two odors. However, since some variance could reflect basic differences in olfactory function (Comfort *et al.*, 1973; Wysocki and Beauchamp, 1984), careful studies of individual differences in discriminative 'odor-spaces' might help scientists understand the coding of quality.

Individual differences in absolute values

Substantial individual differences in overall accuracy and latency also occurred. Differences in speed–accuracy trade-off could play a role here. However, since the slowest subjects did not attain the best performance, nor did the fastest subjects attain the worst (Table 1), it seems individual differences in discriminative capacity may also have played a role. Both factors could influence absolute values of error-based *d*′s. Both factors might also influence absolute values of latency-based *d*′s, though the influence of speed–accuracy trade-off would be less direct. Since latency-based *d*′s represent the difference between two distributions, their values need not differ with overall latency. However, as discussed earlier, speed–accuracy trade-off can influence the slope of the latency versus accuracy function. Differences in bias could also play a role. Bias does not effect error-based *d*′s, but, as discussed earlier, individual differences in degree of bias could cause differences in latency-based *d*′s.

As mentioned earlier, instruction and feedback might reduce differences due to speed–accuracy trade-off or bias. Differences in capacity would remain. Such differences can reflect basic differences in olfactory function (Eskenazi *et al.*, 1986; Martinez *et al.*, 1993), and might deserve further study for this reason. From the standpoint of measurement, however, one must use caution in deriving *d*′s from pooled data when subjects differ in sensitivity or degree of bias (Macmillan and Creelman, 1991). If such differences occur, one can estimate sensitivity for individuals and average the estimates, as was done in the current study, rather than estimate sensitivity from pooled data.

Intended uses

Although discriminative techniques could prove useful in various settings, they might serve best in *quantitative* structure–activity work. Here, the goal is to quantify differences in quality associated with various differences in molecular properties as objectively as possible. We argue elsewhere that discrimination should serve better in this role than any other technique now used (Wise *et al.*, 2000). However, since discrimination specifies the magnitude but not the nature of differences, it could also prove useful to characterize odors via subjective techniques like odor-profiling (Dravnieks, 1985), especially in applied settings. In fact, for some applied work, profiling might provide the information investigators need with less effort.

Conclusions

These studies sought answers to three basic questions regarding response times in olfactory same–different discriminations: (i) does latency reflect accuracy? The data show that latency does reflect accuracy. This demonstrates that latency conveys information regarding qualitative similarity of odors (establishes face validity). (ii) Do measures of similarity based on latency provide better resolution than measures based on errors of discrimination? The data show that *d*′s based on latency do provide better resolution among pairs of odors than *d*′s based on errors. (iii) Do measures of similarity among pairs tested previously predict similarities among pairs not yet tested? The data show that *d*′s based on latency among pairs tested previously do predict *d*′s among pairs not yet tested. Future studies could examine the effects of criterion and speed–accuracy trade- off. For now, though, the results demonstrate that latency to discriminate shows promise as an objective measure of qualitative similarity.

These studies also sought to determine which model best describes the pattern of errors in olfactory same–different discriminations. A model assuming a differencing strategy fit best. In other words, subjects seemed to make relative rather than absolute judgements regarding quality. Future research could determine if subjects use other strategies in different situations. For now, though, the results demonstrate the importance of collecting full ROC curves in order to achieve the best estimate of differences in quality.

Finally, the study sought to examine the qualitative relationship between binary mixtures and their components with an objective measure. The data suggest that the qualities of 50–50 mixtures lie on a straight line connecting the qualities of their components in 'odor-space'. Future studies could continue to use discriminative measures to examine the qualitative relationships between series of proportional mixtures and their components, e.g. to see if additivity holds for mixture ratios other than 50–50. Such studies could provide insights into the coding of mixtures.

Acknowledgements

Supported by research grant 5 R01 DC 00284 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health. The authors would like to thank Dr John Irwin for providing the software used to fit confidence ROCs. The authors would also like to thank Dr Mats Olsson and two anonymous reviewers for valuable comments.

References

Bamber, D. (1969) *Reaction times and error rates for 'same'–'different' judgements of multidimensional stimuli*. Percept. Psychophys., 6, 169–174.

Bindura, D., Williams, J.A. and **Wise, J.S.** (1965) *Judgements of sameness and difference: experiments of decision time*. Science, 150, 1625–1627.

Cain, W.S. (1979) *To know with the nose: keys to odor identification*. Science, 203, 467–470.

Cain, W.S. (1982) *Odor identification by males and females: predictions versus performance*. Chem. Senses, 7, 129–142.

Cain, W.S., de Wijk, R.A., Lulejian, C., Schiet, F. and **See, L.C.** (1998) *Odor identification: perceptual and semantic dimensions*. Chem. Senses, 23, 309–326.

Carterette, E.C., Friedman, M.P. and **Cosmides, R.** (1965) *Reaction-time distributions in the detection of weak signals in noise*. J. Accoust. Soci. Am., 38, 531–542.

Cohen, J.D., MacWhinney, B., Flatt, M. and **Provost, J.** (1993) *PsyScope: a new graphic interactive environment for designing psychology experiments*. Behav. Res. Methods Instrum. Comput., 25, 257–271.

Comfort, A., Whissell-Buechy, D. and **Amoore, J.E.** (1973) *Odour-blindness to musk: simple recessive inheritance*. Nature, 245, 157–158.

Dai, H., Versfeld, N. and **Green, D.** (1996) *The optimum decision rules in the same–different paradigm*. Percept. Psychophys., 58, 1–9.

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

Dravnieks, A. (1985) Atlas of Odor Character Profiles, Vol. 61. American Society for Testing and Materials, Philadelphia, PA.

Emmerich, D.S., Gray, J.L., Watson, C.S. and **Tanis, D.C.** (1972) *Response latency, confidence, and ROCs in auditory signal detection*. Percept. Psychophys., 11, 65–72.

Eskenazi, B., Cain, W.S., Novelly, R.A. and **Mattson, R.** (1986) *Odor perception in temporal lobe epilepsy patients with and without temporal lobectomy*. Neuropsychologia, 24, 553–562.

Espinoza-Varas, B. and **Watson, C.S.** (1994) *Effects of decision criterion on response latencies of binary decisions*. Percept. Psychophys., 55, 190–203.

Francis, M.A. and **Irwin, R.J.** (1995) *Decision strategies and visual-field asymmetries in same–different judgements of word meaning*. Memory Cognit., 23, 301–312.

Francis, M.A. and **Irwin, R.J.** (1997) *Cerebral asymmetry and decision strategies in mental rotation: a psychophysical analysis*. Eur. J. Cogn. Psychol., 9, 225–240.

Gescheider, G.A. (1997) Psychophysics: the Fundamentals, 3rd edn. Lawrence Erlbaum Associates, Mahwah, NJ.

Green, D.M. and **Swets, J.A.** (1966) Signal Detection Theory and Psychophysics. Wiley, New York.

Hautus, M.J., Irwin, R.J. and **Sutherland, S.** (1994) *Relativity of judgements about sound amplitude and the asymmetry of the same–different ROC*. Quart. J. Exp. Psychol., 47A, 1035–1045.

Hummel, T., Hummel, C., Pauli, E. and **Kobal, G.** (1992) *Olfactory discrimination of nicotine-enantiomers by smokers and non-smokers*. Chem. Senses, 17, 13–21.

Irwin, R.J. and **Francis, M.A.** (1995) *Perception of simple and complex visual stimuli: decision strategies and hemispheric differences in same–different judgements*. Perception, 24, 787–809.

Irwin, R.J. and **Hautus, M.J.** (1997) *Likelihood-ratio decision strategy for independent observations in the same–different task: an approximation to the detection-theoretic model*. Percept. Psychophys., 59, 313–316.

Irwin, R.J., Stillman, J.A., Hautus, M.J. and **Huddleston, L.M.** (1993) *Measurement of taste discrimination with the same–different task: a detection-theory analysis*. J. Sens. Stud., 8, 229–239.

Jones, F.N. and **Elliot, D.** (1975) *Individual and substance differences in the discrimination of optical isomers*. Chem. Senses Flav., 1, 317–321.

Katz, L. (1970) *A comparison of type II operating characteristics derived from confidence ratings and from latencies*. Percept. Psychophys., 8, 65–68.

Laing, D.G. and **Willcox, M.E.** (1983) *Perception of components in binary odour mixtures*. Chem. Senses, 7, 249–264.

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.

Laming, D. (1986) Sensory Analysis. Academic Press, London.

Laska, M.L and **Freyer, D.** (1997) *Olfactory discrimination ability for aliphatic esters in squirrel monkeys and humans*. Chem. Senses, 22, 457–465.

Laska, M.L. and **Teubner, R.** (1999) *Olfactory discrimination ability of human subjects for ten pairs of enantiomers*. Chem. Senses, 24, 161–170.

Laska, M.L., Trolp, S. and **Teubner, R.** (1999) *Odor structure–activity relationships compared in human and nonhuman primates*. Behav. Neurosci., 113, 998–1007.

Lawless, H.T. (1989) *Exploration of fragrance categories and ambiguous odors using multidimensional scaling and cluster analysis*. Chem. Senses, 14, 349–360.

Luce, R.D. (1986) Response Times. Oxford University Press, Oxford.

Macmillan, N.A. and **Creelman, C.D.** (1991) Detection Theory: a User's Guide. Cambridge University Press, Cambridge.

Macmillan, N.A., Kaplan, H. L. and **Creelman, C.D.** (1977) *The psychophysics of categorical perception*. Psychol. Rev., 84, 452–471.

MacRae, A.W., Howgate, P. and **Geelhoed, E.** (1990) *Assessing the similarity of odours by sorting and by triadic comparison*. Chem. Senses, 15, 691–699.

MacRae, A.W., Rawcliffe, T., Howgate, P. and **Geelhoed, E.N.** (1992) *Patterns of odour similarity among carbonyls and their mixtures*. Chem. Senses, 17, 119–225.

Martinez, B., Cain, W.S., de Wijk, R.A., Spencer, D.D., Novelly, R. and **Sass, K.J.** (1993) *Olfactory functioning before and after temporal lobe resection for intractable seizures*. Neurophychology, 7, 351–363.

Norman, D.A. and **Wickelgren, W.A.** (1969) *Strength theory of decision rules and latency in short-term memory*. J. Math. Psychol., 6, 192–208.

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

Olsson, M.J. and **Cain, W.S.** (2000) *Psychometrics of odor quality discrimination: method for threshold determination*. Chem. Senses, in press.

Pike, A.R. (1971) *The latencies of correct and incorrect responses in discrimination and detection tasks: their interpretation in terms of a model based on simple counting*. Percept. Psychophys., 9, 455–460.

Proctor, R.W. (1981) *A unified theory for matching task phenomena*. Psychol. Rev., 88, 291–326.

Proctor, R.W. and **Weeks, D.J.** (1989) *Instructional and probability manipulations of bias in multiletter matching*. Percept. Psychophys., 45, 55–56.

Ratcliff, R. and **Hacker, M.J.** (1981) *Speed and accuracy of same and different responses in perceptual matching.* Percept. Psychophys., 30, 303–307.

Ratcliff, R. and **Rouder, J.N.** (1998) *Modeling response times for two-choice decisions*. Psychol. Sci., 9, 347–356.

Sinnott, J.M. (1989) *Detection and discrimination of synthetic English vowels by Old World monkeys*. J. Acoust. Soc. Am., 86, 557–565.

Sinnott, J.M., Brown, C.H., Malik, W.T. and **Kressley, R.A.** (1997) *A multidimensional scaling analysis of vowel discrimination in humans and monkeys*. Percept. Psychophys., 59, 1214–1224.

Smith, P.L. and **Vickers, D.** (1988) *The accumulator model of two-choice discrimination*. J. Math. Psychol., 23, 135–168.

Stevens, S.S. (1950) *Mathematics, measurement, and psychophysics*. In Stevens, S.S. (ed.), Handbook of Experimental Psychology. Wiley, New York, pp. 1–49.

Swets, J.A., Tanner, W.P. and **Birdsall, T.G.** (1961) *Decision processes in perception*. Psychol. Rev., 68, 301–340.

Taylor, D. (1976) *Effect of identity in the multiletter matching task*. J. Exp. Psychol.: Hum. Percept. Perform., 2, 417–428.

Vickers, D. (1980) *Discrimination*. In Welford, A.T. (ed.), Reaction Times. Academic Press, London, pp. 25–72.

Vogals, R. and **Orban, G. A.** (1986) *Decision processes in visual discrimination of line orientation*. J. Exp. Psychol.: Hum. Percept. Perform., 12, 115–132.

Welford, A.T. (1960) *The measurement of sensory-motor performance: survey and re-appraisal of twelve years' progress*. Ergonomics, 3, 189–230.

Wise, P.M., Olsson, M.J. and **Cain, W.S.** (2000) *Quantification of odor quality*. Chem. Senses, in press.

Woodworth, R.S. and **Schlosberg, H.** (1954) Experimental Psychology, revised edn. Holt, Rinehart & Winston, New York.

Wysocki, C.J. and **Beauchamp, G. K.** (1984) *Ability to smell androstenone is genetically determined*. Proc. Natl Acad. Sci. USA, 81, 4899–4902.

Accepted December 1, 1999