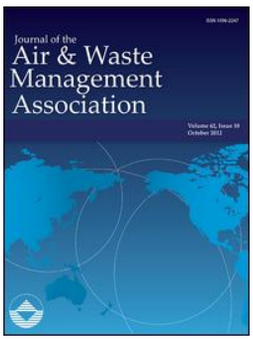

Journal of the Air & Waste Management Association

ISSN: 1096-2247 (Print) 2162-2906 (Online) Journal homepage: https://www.tandfonline.com/loi/uawm20

A Review of 20 Years of Standardization of Odor Concentration Measurement by Dynamic Olfactometry in Europe

A.P. (Ton) van Harreveld , Paul Heeres & Hendrik Harssema

To cite this article: A.P. (Ton) van Harreveld , Paul Heeres & Hendrik Harssema (1999) A Review of 20 Years of Standardization of Odor Concentration Measurement by Dynamic Olfactometry in Europe, Journal of the Air & Waste Management Association, 49:6, 705-715, DOI: 10.1080/10473289.1999.11499900

To link to this article: https://doi.org/10.1080/10473289.1999.11499900

TECHNICAL PAPER

ISSN 1047-3289 *J. Air & Waste Manage. Assoc.* **49**:705-715

A Review of 20 Years of Standardization of Odor Concentration Measurement by Dynamic Olfactometry in Europe

A.P. (Ton) van Harreveld
Project Research Environmental Consultants Ltd., Avoncliff, Wiltshire, United Kingdom

Paul Heeres and Hendrik Harssema
Agricultural University of Wageningen, Department of Air Quality, Wageningen, the Netherlands

ABSTRACT

Twenty years of development of standards for odor concentration measurement in Europe are reviewed. Issues covered are the development of instruments, standards, and guidelines for methodology and calculation of odor concentration as well as validation of results through interlaboratory comparisons. National efforts led to an initiative in the technical committee CEN/TC264 Air Quality of the Committee Européen de Normalisation (CEN) to draft a European standard. The final draft was validated by an international interlaboratory comparison in 1996. This validation showed that the strict performance criteria set for repeatability, reproducibility, and accuracy (assessed for an n-butanol standard) were attainable by laboratories using two modes of presenting odor samples to assessors (single-stimulus and forced-choice) that were included in the standard. An agreed-upon reference value for the European odor unit was set at 1 $ou_E/m^3 \equiv$ 40 ppb/v n-butanol. This value is attained by a strict protocol for panel selection, which has been the main contributing factor in improving the reproducibility of olfactometry. The notion that the panel should be representative of the general population was explicitly abandoned, recognizing that this proposition was not tenable using small panels of a practical size of five to eight assessors. The validation demonstrated that the CEN draft method, expected to be introduced in 1999, can guarantee reliable and reproducible results, which are effective in supporting the implementation of effective odor abatement policies in the member countries of the European Union.

IMPLICATIONS

A sizable proportion of environmental complaints are raised due to odors. Environmental odor management requires tools to quantify, regulate, and enforce odor emission limits. The application of olfactometry has been hampered by a lack of reproducibility and traceability of results. The introduction of air quality guidelines for odors in Europe in the mid 1980s has prompted further standardization of olfactometry and validation through interlaboratory comparisons. The most recent development is the drafting of a standard for dynamic olfactometry by the European standardization organization CEN (in Brussels). This standard was validated in a large scale interlaboratory comparison in 1996 and is expected to become mandatory in the European Union in 1998.

INTRODUCTION

In the 1970s, North European countries and the United States alike developed a framework of environmental legislation to cope with the increasing pollution of air, water, and soil. The air quality programs were mostly aimed at reducing SO_2, smoke, and other chemical compounds from combustion and stack exhausts. It was not until the late 1970s and early 1980s that attention became focused on odor annoyance problems.[1] For the community, odor pollution has been a common nuisance and cause of complaints.[2-6] The abatement of odor pollution was, in most cases, left to local authorities generally using personal observations and judgment to assess the magnitude of nuisance. This practice was open to error, subjectiveness, and bias, and therefore was not very satisfying for the community, authorities and owners/operators of the sources. The need for an adequate, objective method to measure odor concentration to characterize emissions became obvious, prompting development of practical methods and standards for olfactometry.

The measurement of odor was not a scientific novelty, however. Physiologists interested in the functioning of the olfactory sense[7] have been involved in olfactometry since the late nineteenth century. This field of research yielded the methodology to measure sensory thresholds as a physiological parameter of scientific interest. The issue of odor detection thresholds and their variability between individuals and within one individual over time was a research subject that was generally addressed by statistics.[8-10] The results of research on

properties such as "hedonic tone," "intensity," and "odor quality" found their application in the food and flavor industry.

The application of olfactometry as a *practical* tool for environmental application raised the important question how large the panel had to be to act as a representative sample for the general population to assess the characteristics of response to odors.[11-15] Unfortunately, the gap between performance of a practical, economic measurement and the needs for an objective and accurate instrument to be used in odor abatement policy remained unacceptably wide. Odor thresholds for pure substances published by different researchers remained notoriously large, spanning several orders of magnitude.[16]

Implementation of odor policy, abatement programs, and regulations, however, requires a reliable and accurate measurement with acceptable repeatability, accuracy, and traceability to provide valid odor characterization data that remain stable in time. In the mid 1980s, the reduction of odor annoyance became an important issue in northern European countries. In some cases, a strict regulatory approach, based on air quality guidelines, was introduced.[17] As a result, the need for suitable tools for odor measurement became urgent, leading to development of recommendations, guidelines or (draft) national standards in countries such as Sweden,[18] Denmark,[19] United Kingdom,[20] Germany,[21] the Netherlands,[22] and France.[23]

To improve and validate such standards, interlaboratory comparisons were conducted in Germany [13,24-30] and the Netherlands.[31-35] When the results of validating the Dutch pre-standard[36] indicated that the required quality for environmental application was within reach, a working group was formed in the framework of the Committee Européen de Normalisation (CEN) to draft a European olfactometry standard. This working group, CEN\TC264\WG2 "Odours," had the benefit of the active participation of experts from ten countries and produced a final draft in 1996. To validate the CEN draft standard, a large-scale international interlaboratory comparison using n-butanol was conducted in 1996.[37]

This paper reviews these developments, the results of the interlaboratory comparisons, and validation of the European draft CEN standard based on an assigned value for the odor threshold of a reference odor, including strict performance criteria for repeatability, reproducibility, and accuracy.

DEVELOPMENT OF OLFACTOMETRIC METHODS

When the need for practical olfactometry for environmental applications arose in the 1970s, olfactometers were not commercially available. Institutes, agencies, and environmental consultants with an interest in odors designed and built their own olfactometers. In the United States, the olfactometers built by Hemeon[38] and Dravnieks[39] were examples of this interest. The ASTM D-1391[40] guideline described a dilution method using syringes. The ASTM standard practice E679-91[41] described a protocol for forced-choice olfactometry. In Europe, there was an even wider variety of approaches to odor assessment. Researchers in countries such as Finland,[42] Sweden,[43] Denmark,[44] the United Kingdom,[45] the Netherlands,[46,47] Germany,[48,49] Belgium,[50] France,[51] and Switzerland[52] developed all sorts of dilution devices and odor assessment methods. At that time there were no specific or standardized requirements, for instance, for the performance of the olfactometry instruments. As long as the application of olfactometry results in the environmental management of odor was still in its infancy, there was a wide range of acceptance for odor assessment protocols. Although results were shared between olfactometrists, the considerable differences between laboratories were accepted as a matter of fact. Toward the early 1980s, the interest in reducing these differences grew, and a limited number of comparison studies between different olfactometers and methods were initiated. Dravnieks and Prokop[53] compared the performance of their newly developed "forced-choice" triangle, a dynamic olfactometer with the ASTM syringe method, and the olfactometer developed by Hemeon.[38] Thiele et al.[13] in Germany and Schaefer[54] in the Netherlands compared the performance of different types of olfactometers.

Instrument Standardization

Based on such comparisons, Dravnieks and Jarke[12] analyzed the most relevant operational parameters for odor assessment and proposed a series of measures to improve the reproducibility of results between laboratories. Their remedy was to standardize the dilution equipment (i.e., the olfactometer) and the assessment protocol because their results showed considerable differences between results of olfactometers of different designs. The ratio between lowest and highest thresholds measured on one single odorant using different olfactometers was about 200, while this ratio was about 2.5 for the results obtained by one dynamic "state-of-the-art" olfactometer. For instrument standardization, Dravnieks recommended

(1) the use of a dynamic dilution device with a fixed, increased air flow to the sniffing ports (e.g., 9 to 200 L/min);
(2) two or three sniffing ports of a well-defined design;
(3) the use of a fixed or standard series of dilutions; and
(4) the use of clean, odorless dilution air.

For standardization of assessment methodology, Dravniek recommended

(1) the choice mode of "forced-choice" (choice between stimulus and one or more odorless ports) over the yes/no (single-stimulus) mode;

(2) presentation of stimuli (i.e., odor samples) in an ascending series of intensities (or decreasing dilution factors);
(3) selection of assessors, with exclusion of highly sensitive or insensitive individuals (Note: Dravnieks noted that that reproducibility would be improved by selecting assessors with the same sensitivity but warned that this could bias the representativeness of the result for the entire population.);
(4) a number of at least eight assessors to ensure proper data validation, or at least two presentation series to each assessor for panels of five assessors;
(5) ensuring circumstances for independent and unbiased decision of the assessor; and
(6) use of the geometric mean of individual threshold values (for one assessor in a panel) to calculate the panel threshold.

At this early stage, some parameters were already recognized as being of great importance for the outcome of the olfactometric analysis. The flow and velocity of odorous air emanating from the ports was identified as a significant influence on the outcome of the odor threshold.[53] The Hemeon olfactometer detected much lower (by a factor of 6) thresholds with its large volume flow of air (150 L/min) compared to that of Dravnieks (0.5 L/min). Using this approach and instrument, Dravnieks intended to produce results comparable with results generated by the then-existing ASTM syringe technique. He succeeded in that he noted a remarkable similarity in results. The approach in Europe regarding sample flow and velocity was that the olfactometer should provide the assessor with sufficient air to be compatible with normal breathing and sniffing rates, avoiding extra dilution through inhalation of "room air."[22,55-57] In his study of sniffing rates in humans, Laing[55] proposed an air flow of 40 L/min. In the Netherlands, a simulation study using tracer gases and an anatomical model of a nose[56] showed that at flow rates of 20 L/min, there was only a minor decrease of the threshold value compared to that determined at higher flows. The port design was such that the nose could be held in the port. This study concluded that air velocity should be about 6 cm/sec as a minimum. A recent U.S. study[58] showed similar trends for the relation between flow rate and odor threshold.

Perhaps the most widely mentioned parameter to determine outcome of results is assembling a suitable panel of assessors for odor assessment (olfactometry). In physiological research, panels of 50 or more assessors were used to study olfactory characteristics of the general population.[59] For environmental odor assessment, this approach was impractical and too costly. Although researchers generally supported the goal of panels being representative of the general population, the number of assessors was reduced sharply, from approximately 50 to 5, as a practical consideration.[12] This exacerbated the issue of panel selection, since the variability in olfactory sensitivity between individuals and within an individual in time is considerable.[11,60] When using small population samples, it becomes increasingly important to select assessors on the basis of knowledge about their olfactory sensitivity to certain odorants. After all, in odor research, the threshold is usually defined as the dose that 50% of a population can detect as a sensory stimulus (D_{50}). However, how to obtain a threshold for the general population from a limited number of individual assessor thresholds obtained in the assessment procedure remained a problem. Finney,[8] Drake,[9] and Dravieks et al.[10] developed statistically based methods to calculate the D_{50} detection threshold of a panel. The panel under survey, however, consisted typically of 8 to 16 individuals. As a random sample, this poorly represents the population as a whole. As a result, considerable variability between results obtained by different panels was observed.[12,13,15] To reduce this variability, the general approach was to exclude highly "sensitive" or "insensitive" individuals from the tests. In Germany, Thiele et al.[13] tested a group of assessors to gain more insight into this subject. They started with a group of 35 panel members and excluded individuals at the extremes of the frequency distribution of individual threshold values. They excluded all but 15 assessors to conduct their research.

In the early 1980s, the increasing application of olfactometry for environmental policy implementation and licensing of industry highlighted the need for improved reproducibility of results.[17] As a result, national (draft) standards or guidelines were adopted in Denmark[19], United Kingdom,[20] Germany,[21] the Netherlands,[22] and France.[23] To test these (draft) standards or guidelines, a number of interlaboratory comparisons (ILCs) were conducted, first in Germany in the first half of the 1980s, then in the Netherlands in the latter part of that decade, and lately in the European Union in 1996. All interlaboratory comparisons were designed to assess whether the formulated criteria would lead to an acceptable level of accuracy[61] and support an adequate and effective odor abatement policy.

German Interlaboratory Comparisons (Ringversuche)

Thiele et al.,[13,24,25,27,28] Bahnmüller,[26] Dollnick et al.,[29] and Both et al.[30] conducted a series of ILCs with participation from German laboratories. The general goal of the ILCs was to provide answers to the following questions:

- What is the repeatability of the odor measurement results? In contrast to the concept of repeatability as used in analytical laboratories, the Germans distinguished between a series of olfactometric analyses during one day, to determine

short-term repeatability, and a series of analyses dispersed over weeks or months, to determine long-term repeatability.
- What is the reproducibility of the results of olfactometric analyses in two or more laboratories on one or more odor substances?
- Are the performance characteristics measured on the test odor substance transferable to other odors or mixture of odors?
- Are the results of individual assessors in a panel representative of the population under survey, and what are the characteristics and distribution of the responses of the assessors used in the measurements?

On the basis of the ILC held in 1981, Thiele et al.[13] reported the following remarkable conclusions:

- The repeatability measured using the same panel and equipment is the same within one day as it is within one year, that is, there was no difference in short- versus long-term repeatability. The variance was larger when replacement panel members were used.
- The repeatability of the results improved with the number of replicates within one measurement up to five replicates. In practice, three replicates are the optimum, considering other aspects of the assessment such as cost.
- The reproducibility was mainly a function of the actual dilution equipment used, the presentation mode (the experimental psychophysical conditions), and the variability within one assessor and between individual assessors.

With these conclusions, they recommended improving the reproducibility of odor assessment through development of a "standard panel" calibrated on a standard test substance such as H_2S. They stipulated that such a standardization is conditional for the results being transferable to other odorants, and that this should be tested by determining whether the repeatability and reproducibility measured by such a panel for other odors are compatible with those for the test substance.

The ILC of 1982 reported by Thiele[25] was conducted to answer the following questions:

- What is the reproducibility of odor assessment results obtained for a test substance by laboratories using varying equipment, protocols, and panels?
- Are instrument characteristics (e.g., type of sniffing port and air flow rate) significant variables affecting reproducibility?
- What is the effect on the odor assessment results associated with the selection of panel members at the different laboratories?
- Are results transferable to other substances?

In his conclusions, Thiele noted the effect on the reproducibility of odor measurement results of (1) the air flow rate emanating from the ports, (2) the variability caused by using different panels, and (3) the variability of individual results within a panel. He identified age as a factor and proposed to exclude persons above 50 from participating in a panel because of their diminishing sensitivity to smell.

The issue of transferability was not discussed because he only used H_2S as a test substance.

The ILC of Thiele[27] in 1984 investigated the degree to which the performance characteristics measured for H_2S are transferable to an environmental test odor. To answer this question, the participants analyzed both an environmental odor and the reference odor substance (H_2S). His objective was to determine whether the performance characteristics for the environmental odor were compatible with those found for reference odor results (H_2S). To assess transferability he used the standard deviation of the odor measurement results for H_2S compared with that calculated for the environmental odor as the characteristic variable. He also investigated the effect of instrument differences in detail.

His conclusions supported these earlier findings:

- The flow of air (with or without odor sample) presented to assessors should be sufficient for normal breathing.
- The forced-choice presentation mode yields lower thresholds (in concentration terms) than the yes/no mode.
- None of the other instrumental characteristics considered were important.
- The assessor age is a factor influencing results.
- The overall results, corrected for outliers and other apparent faults, showed a reproducibility of a factor of 5. This factor indicates that the difference between two single test results, obtained by different laboratories, is less than a factor of 5 in 95% of cases.

The ILC reported by Bahnmüller[26] in 1984 focused on the issue of transferability. He selected five odorants, including H_2S, dibutylamine, acrylic acid methyl ester, isoamyl alcohol, and a common mixture of solvents for metal coating purposes. He analyzed his results by calculating the standard deviation of the logarithms of the measured odor threshold results (in $\mu g/m^3$). Then he calculated the quotient of the 84- and 16-percentile values of the distribution by calculating the antilog of 2 times the standard deviation. He observed that this quotient, for the different odorants tested, did not differ significantly (i.e., within a factor 2 to 3). Using this indicator, he demonstrated that the performance of the odor assessment was transferable. However,

note that the test characteristic as defined by Bahnmüller is a quotient comprising only (2 times) 1 standard deviation (16 and 84 percentiles), which excludes the impact of extreme values from the assessment. For instance, if the quotient is calculated based on an interval of (2 times) 2 standard deviations, the results would have shown greater differences, of a factor 4 to 45.

In 1985 Thiele and Bahnmüller[28] reviewed the results of their earlier work. They came to the conclusion that the identification and exclusion of "outliers" in odor research is often arbitrary because of the considerable variability in the measurement results. A rather disappointing conclusion was that the use of H_2S as a reference odor did not significantly improve the results of the ILCs.

In 1988, Dollnick et al.[29] published the results of an ILC conducted in 1985. Again five substances were involved: H_2S, n-butanol, isoamylalcohol, propionic acid, and dibutylamine. The objective was to answer questions such as

- What is the reproducibility between laboratories, compared with earlier ILC results when using a standard odor sample preparation and delivery procedure for the test odorants?
- Are the performance characteristics of the results transferable for the test odorants (assessed on the basis of the quotient of the 84 and 16 percentiles)?
- Do differences in instrumental characteristics such as air flow from the ports and presentation decision mode (i.e., forced-choice or yes/no) affect the odor assessment results?

Their conclusions did not indicate a significant improvement over earlier ILCs. The reproducibility of results was even less favorable than in earlier ILCs. After stratification for instruments and presentation mode, the reproducibility was the same as in earlier ILCs. He applied the same quotient approach as in the earlier work of Bahnmüller to evaluate transferability, applied to results of all participants and all odor assessment protocols. They reported a quotient between 10 and 30. After stratification for instrument type and presentation mode, this quotient was 5 to 8. This led Dollnick et al. to conclude that the transferability was clearly demonstrated. As with the ILC conducted earlier by Bahnmüller, the comment can be made that when the quotient is calculated for an interval of (2 times) 2 standard deviations, the quotient varies between 100 and 200 for the different test substances.

To achieve better performance in terms of reproducibility they recommended that the method for odor assessment be standardized for parameters demonstrated to influence results, and that further ILCs be organized on a regular basis, involving at least three reference odors, to provide laboratories with the opportunity to perform routine performance evaluations and to modify their procedures.

The latest in this series of German ILCs was conducted by Both et al.[30] in 1992. They arranged for all measurements in the ILC to be carried out at one location, which allowed all four participating olfactometry teams to use the same panel. To assess repeatability and reproducibility (ISO5725[61]), their test design involved two test substances (n-butanol and dibutylamine) on three concentration levels, with three replicate measurements at each level. In the test results, they applied the concept of reproducibility and repeatability, calculated according to ISO 5725. Repeatability and reproducibility were calculated from the logarithms of the measured thresholds (in $\mu g/m^3$). Then the antilog of the repeatability and reproducibility were calculated, representing the maximum factor by which two single measurements will be apart in 95% of cases. These test results showed a repeatability of a factor 2 to 5 for the different test substances and a factor 5 for the reproducibility. Although these results are considerably better than results from earlier ILCs, test conditions were restricting in that testing was confined to one panel and a limited number of operators working in one location. In their recommendations, Both et al. proposed

(1) stricter standardization of the odor assessment protocol, as compared to the existing standard VDI 3881;
(2) additional ILCs with
   (a) replicate presentations on each concentration level,
   (b) checks on the actual test odor concentrations using analytical methods,
   (c) performance evaluation of the olfactometer dilution characteristics, using tracer gas and analytical methods (e.g., carbon monoxide with NDIR detection), and
   (d) participation of a wider variety of olfactometers and operators in the ILC; and
(3) defining the values of the quality parameters for accuracy (including reproducibility and repeatability) using reference odors.

On the basis of the results of all the ILCs mentioned, the standard VDI 3881 was updated in 1989. The concepts of repeatability and reproducibility, as described in ISO 5725 were adopted. These concepts were applied to the logarithm of the odor assessment results. Also, Supplement 4 was added to the standard document VDI 3881,[21] summarizing results of the ILCs and their conclusions:

(1) The sensitivity to odorants is age dependent, and assessors are limited to the ages of 18 to 50 years.
(2) The forced-choice mode yields lower thresholds (in concentration terms) than the yes/no mode.
(3) Inadequate air flow from the ports yields high thresholds (in concentration terms).

(4) Presentation of odor stimuli in ascending order yields lower thresholds (in concentration terms) than the presentation of odor stimuli at random.
(5) The repeatability measured on the odor test substance, H_2S, should be between a factor 3 to 5 for long-term and a factor 2 for short-term repeatability. The reproducibility should be within a factor of 10 to 190 for all instruments and methods and from 10 to 40 within a method and instrument.

The VDI 3881 (1989) also recommended that (1) a control measurement be performed with a reference odor for each measurement session, (2) repeatability should be within a factor of 3, and (3) reproducibility measured on n-butanol or H_2S be within a factor of 25 and lie within the following ranges:

H_2S: 0.6< measurement result< 15 $\mu g/m^{-3}$
Butanol: 110 < measurement result < 2800 $\mu g/m^{-3}$

It was also suggested that the overall quality check of the measurement results be extended to the individual thresholds of each assessor, applying the same criteria for repeatability, on their individual performance.

Dutch Interlaboratory Comparisons (Ringtesten)

In the Netherlands, the standardization process was "kick-started" by the introduction of an odor policy requiring olfactometric data by the Ministry of Public Planning and the Environment in 1985.[17] A first step toward standardization was found in the recommendations for instrument parameters and protocol summarized by Roos.[22] These included

- the air flow from the ports should be at least 20 L/min;
- the ports should be of well-defined design;
- the forced-choice presentation mode should be used;
- a minimum of eight assessors should be used;
- each presentation series should be replicated twice in each measurement;
- a minimum of four dilution steps should be used in each presentation series, with a fixed step factor between 1.4 and 3;
- the protocol should include fixed sniffing and resting periods;
- the instrument calibration should be performed using a reference gas;
- ascending concentration in presentations should be used;
- one calculation method should be defined in the standard protocol;
- assessors should be selected based on sensitivity for odorants;
- testing should be conducted in an odorless environment; and
- good working conditions (mental and physical) should be maintained in the odor room.

The first Dutch ILC was organized in 1986 and reported by Hermans[31] in 1989 and by Heeres and Harssema[32] in 1990. The results were similar to those in Germany, which were not satisfactory for implementation of the odor policy in force at the time. The reproducibility was about a factor of 20 for the three odorants ethylbutyrate, butanol, and H_2S. (The same procedures for calculating repeatability and reproducibility were used as described earlier in the work of Both et al., again using the concepts of ISO5725.) The results showed similar performance for all three odorants. In 1987, the working group of olfactometry experts, WG 390 146 01 15 "Olfactometrie," assembled by the Netherlands Normalization Institute, started drafting a national pre-standard (NVN2820[36]), which was a performance standard based on the concept of ISO5725. The quality requirement for policy implementation was defined as a reproducibility better than a factor 4 and a repeatability better than a factor 2.

To determine whether standardization of protocol achieved this goal, a long-term ILC was designed and organized to operate over a three-year period. Certified standard reference materials for n-butanol and H_2S at five concentration levels were developed and distributed to participating laboratories. Testing by 10 participating laboratories was conducted monthly on four samples. Furthermore, all participants had to comply with the instrumental and methodical requirements listed earlier. A calibration procedure for the dilution performance of olfactometers was developed, and all olfactometers had to comply with performance characteristics for the dilutions produced. No specific requirements for panel member selection were formulated at that stage.

The Dutch long-term ILC started in 1990 and ended in 1993.[33-35] The first-year results showed that the quality objective for repeatability was not achieved, and reported repeatability for individual laboratories with a range of factors from 3 to 20, with an extreme factor of up to 300 reported in an isolated case. Following the initial results, a performance analysis was made of individual assessors in the panels.[33] These results showed a repeatability factor of 3 to 5 within the performance of an assessor and a factor up to 50 (or greater in some cases) between assessors. To reduce this variance, which seemed to be caused in part by the forced-choice mode, the assessors were asked to make an additional assessment at each odor sample presentation and to indicate whether their response was given on the basis of guessing, by having an inkling, or with certainty. This "enhanced choice" mode reduced the variance of results within an individual but did not significantly improve the repeatability of overall panel threshold results. As a result, strict

requirements were then formulated, and assessors were selected on the basis of compliance to these assessor performance criteria. Based on an assessment of at least 12 individual thresholds for n-butanol collected and measured over several days, the thresholds of the individual assessor had to fall within a factor 2 of the geometric mean value of the panel results. In addition, the antilog of the standard deviation of the logarithms of the thresholds had to be less then a factor 2.5 of the measured geometric mean for that assessor (in ppb n-butanol). These measures led to a general improvement of repeatability to a factor of 5 to 10, with an extreme of 80 in an isolated case. Still, the results—a factor of 2 or less—were far removed from the target for repeatability. To effectively reduce the variance between assessors, an extra criterion was added: The individual threshold on n-butanol, using the forced-choice-with-certainty mode, must fall within the range of 20 to 80 ppb/v. This strict selection of assessors finally produced the required result. The repeatability was reduced to a factor varying between 1.5 and 3, with a factor of 5 in an isolated case. In a final ILC[35] conducted in the winter of 1993 under the auspices of the Netherlands Calibration Organization, the ultimate round of testing took place. Each laboratory analyzed three concentration levels of n-butanol, quadruples of each concentration level. The sample identification was randomized, giving no indication of the content of the sample. The results showed a reproducibility factor of 3.3 and a repeatability varying from 2 to 3, with an extreme of 4.5. Results from this ILC were the basis of the quality criteria set in the final Dutch Standard NVN2820, issued in 1995.

The critical difference between the Dutch and German standards is the definition of quantitative performance criteria for the selection of assessors, performance criteria for the olfactometer instrument, and overall performance criteria for each laboratory, assessed for a standard reference material n-butanol. Figure 1 illustrates a typical result of a German ILC and the progression of results in consecutive Dutch ILCs. It shows clearly the improvement achieved in March 1993 by improved panel selection and use of an accepted reference value of 20 ppb/v n-butanol. The shift from an accepted reference value of 20 ppb/v initially agreed upon in the Netherlands (1992 to 1993) to the consensus value agreed upon in the CEN draft (ICO 1996) to 40 ppb/v is also evident.

CEN International Comparison of Olfactometry (ICO)

In 1990, CEN assembled working group CEN/TC264/WG2 "Odors" to standardize the olfactometric measurement method and draft a European standard.[62] Experts from Finland, Denmark, Ireland, United Kingdom, the

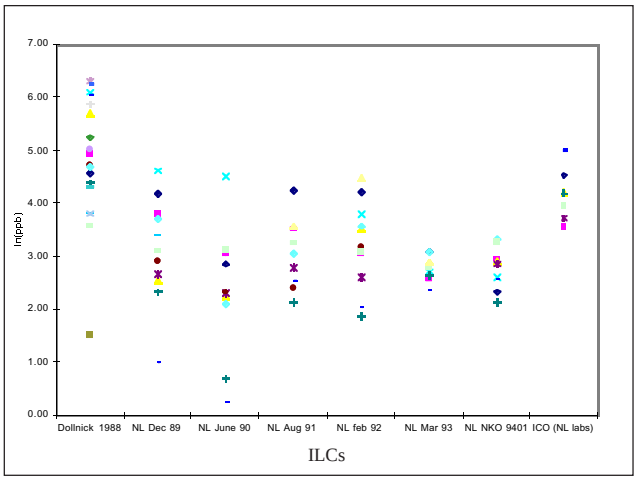

Figure 1. Mean threshold values for n-butanol, in ln(ppb/v), measured by laboratories participating in interlaboratory testing in Germany and the Netherlands, 1988–1996.

Netherlands, Germany, Belgium, France, Switzerland, and Austria actively contributed. In drafting the standard, maximum flexibility was allowed both in the prescribed protocol and in requirements for equipment, including construction of instruments, presentation mode, and order of presented dilution levels. A minimum number of panel members (four) and replicate presentation series (two) was prescribed after an analysis of a large set of raw data on replicate odor measurements by Harreveld and Heeres.[63] The core of the standard, however, is a set of strict performance criteria for the dilution instrument, the selection and performance of assessors, and the olfactometric measurement procedure as a whole. The unit of measurement was defined and made traceable to an accepted reference value of a reference material:

- A *European odor unit* (ou_E) is that amount of odorant(s) that, when evaporated into 1 m^3 neutral gas at standard conditions, elicits a physiological response from a panel (detection threshold) equivalent to that elicited by 1 European reference odor mass (EROM) evaporated in 1 m^3 neutral gas at standard conditions.[62]
- The *European reference odor mass* is the accepted reference value for the European odor unit, equal to a defined mass of a certified reference material. One EROM is equivalent to 123 µg n-butanol (CAS 71-36-3). Evaporated in 1 m^3 neutral gas, this produces a concentration of 0.040 µmol/mol.[62]

The reference material n-butanol is used as the basis for a quality control and assessment structure, including regular performance evaluation of assessors and of the measurement procedure (system audit).[62] The most important performance criteria are summarized below.

- Calibration of the dilution equipment (olfactometer) using certified tracer gas standards to assess the accuracy criterion: The actual dilution factor should be within 20% of the expected value, at a

confidence level of 95%. The instrumental repeatability is implicitly a part of this criterion.

- Selection of assessors based on their individual sensitivities and variabilities in the detection of certified n-butanol: The antilog of the standard deviation of at least 12 individual threshold estimates (ITEs), expressed as log(ppb/v), should be less than a factor of 2.3, while the geometric mean should be between 20 and 80 ppb/v.
- Overall quality criterion of accuracy of group threshold results for n-butanol: The accuracy, A_{od}, must be less than 0.217 (calculated from the logarithms of test results expressed as threshold in ppb/v). This criterion can be expressed as a factor of 1.6 when applying the antilog. The accuracy is a combined measure of how far the geometric mean value of results for odor concentration is removed from the accepted reference value and the variability of these results. If the geometric mean value of the thresholds is 40 ppb/v, the accuracy criterion implies that 95% of results must be in the range of 25 to 64 ppb/v. If the mean is further removed from the 40 ppb/v target, the range of random variability must be narrower than mentioned above to comply with the accuracy criterion. Accuracy is assessed on the basis of at least 10 panel threshold values, collected over a period of less than 12 months.
- Repeatability criterion for group threshold results for n-butanol: The repeatability, *r,* must be less than 0.477 (calculated from the logarithms of test results expressed as threshold in ppb/v). This criterion can be expressed as a factor of 3 when applying the antilog. This implies that 2 single panel threshold results may not be more than a factor 3 apart in 95% of cases. This is also assessed on at least 10 panel threshold results, collected over a period of less than 12 months.

The procedures of the CEN require validation of the draft standard through interlaboratory comparison to confirm that set performance criteria are attainable. To do so, the International Comparison of Olfactometry (ICO) was organized in 1996, with 19 laboratories from five countries participating. The results of the ICO confirmed that the overall criteria of the draft standard were attainable.[37] The draft will be published in 1998 for inquiry, followed by voting on the adoption of the draft as a European Standard (EN) by the CEN members later that year. The standard is expected to be adopted and implemented toward the end of 1998, replacing the national standards that are now in use in the European member states.

DISCUSSION

The issue of transferability of performance on different substances was recognized early in the development of odor measurement.[10,11,27,29,59,60,64] To ensure that olfactometry can be generally applied to the assessment of samples of environmental odorants, it must be confirmed that the selection of assessors for odor panels, on the basis of their performance on the reference odor, is also predictive for their performance on environmental odorants. Quality criteria for the reference odor can be considered to be *transferable* if the variability for those environmental odorants is similar to that for the reference odor.[63,64] The German ILCs from 1979 to 1988 and the Dutch ILC of 1985 concluded that transferability is demonstrated for a number of pure substances within their criteria, which are not universally accepted. Punter,[59] however, showed a different sensitivity of assessors for different *pure* substances. Laska and Hudson[64] investigated pure substances and multi-component mixtures of these same pure substances and observed that the variability in personal thresholds measured for the single substances was generally larger than for the mixtures. Harreveld and Heeres[63] analyzed a number of butanol and environmental odor panel thresholds. They compared the standard deviations of the frequency distribution of the panel thresholds for both odors and concluded that variability was least for the environmental odors. The same effect was reported in the CEN ICO study of 1996, in which, as an additional program, 7 laboratories each analyzed 10 samples of an environmental odor sample (a yeasty odor) at 5 concentration levels.[37] In Figures 2a and 2b, the mean results (n=10) for those laboratories that analyzed both (a) n-butanol reference and (b) environmental odor in the ICO are plotted against expected values. The figures illustrate that the variability for the environmental odor is less than for n-butanol. Considering that these measurements were conducted within the strict criteria of reproducibility and repeatability,[61] the evidence of transferability is much stronger than that obtained in earlier results. This work supports the conclusion that quality parameters assessed for a reference odor are transferable to environmental odors. The CEN working group, however, recognizes that a reference mixture of 5–10 odorants would be preferable as a reference material. Attempts are underway to develop such a multi-component mixture as a reference material.

Another contentious subject is the question of whether the panel used can be, or should be, considered representative of the population under survey.[11-15] Researchers agreed that from a sample of assessors the most sensitive and insensitive assessors should be excluded from a panel. The

remaining assessors could be included in the panel and perform measurements, with a suggested panel size of about 20–30 assessors for basic research, 8–12 for daily routine measurements, and 4–6 for comparative measurements. They suggested excluding the most and least sensitive assessors after comparing individual thresholds for pure odorants to the panel threshold. This selection process, however, was not quantified in the protocol and was left to the discretion of the operator. Subsequent standards [23,36,62] provided strict statistical criteria for panel selection based on a minimum set of data collected for each assessor. To provide a further check on transferability that can be applied to each measurement, Harreveld and Heeres[63] developed a quantitative criterion to exclude assessors with "deviant" results for the odor sample being tested on the basis of a retrospective screening. This criterion is included in the procedures of the draft CEN standard.[62] Each of these selection and screening steps, however, takes us further away from having a representative sample of the population. While the most important improvement in reproducibility was achieved by selection of assessors with similar sensitivities to form a panel, it also implies selection of a subset from the population. In drafting the CEN standard, the notion that the panel should be representative of the population was explicitly abandoned. Instead, a well-defined subset of the population is to be selected. In order to define the selection criteria, the CEN working group collected all available thresholds measured in the laboratories of its members on butanol (the proposed reference odor). As expected, a wide range of values were reported, but a consensus was reached that 40 ppb/v was a value that all laboratories could attain after implementing the proposed methodology. Most laboratories can attain the 40 ppb/v target by selecting between 30 and 50% of the recruited novice assessors. Although we know that the 40 ppb/v anchor point apparently falls within the distribution of sensitivity to n-butanol for the entire population, it is not known how close this value is to the median value. To determine the frequency distribution of threshold values for n-butanol and its median, a genuinely representative sample of the population would need to be surveyed. A representative sample of the population can be obtained as follows: (1) The population must be accurately defined and delimited; (2) the sample drawn must be truly random, with every member

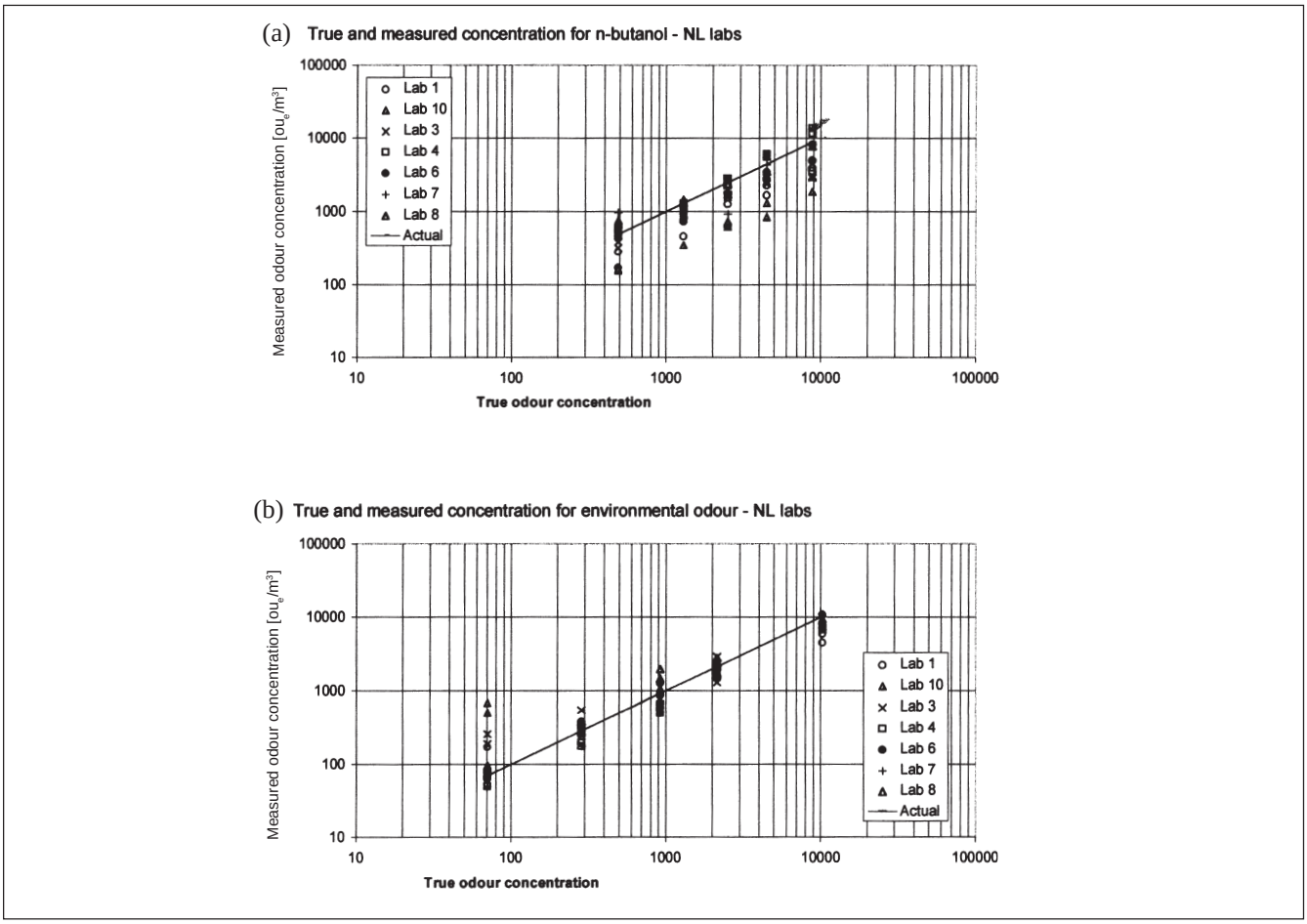

Figure 2. True concentrations in European odor units (ou_E/m_3) plotted against measured concentrations for n-butanol and an environmental odor, Dutch participants in the CEN-ICO 1996.

of the population having a known chance of being selected; and (3) knowledge of the degree of variation within the population exists or can be acquired in the course of formulating the plan of sampling.

Following these requirements, it is clear that the sample size may be several hundreds of individuals, which is obviously impractical for use in a practical testing method on a daily basis. It would be useful, however, to determine the position of the consensus value of 40 ppb/v n-butanol in the frequency distribution for the population. This would be a one-time test giving results that could be used by all.

CONCLUSION

After reviewing 20 years of development of olfactometry in Europe, the main body of scientific literature suggests that an effective odor abatement policy is attainable, but only after implementation of strict standardization of measurement protocol and quality control procedures. The major breakthrough in reducing the variability in results and achieving convergence in results obtained by different laboratories was achieved only after taking two crucial steps:

(1) Panel selection: Strict control of the "sensor" characteristics must be attained by selecting assessors with similar sensitivities, whose responses are consistent in time. Assessors are selected on the basis of applying statistical criteria to their responses for a reference odorant.

(2) "Span adjustment": Once an agreed-upon reference value for the odor unit was defined for a reference material (n-butanol), laboratories could adjust their method to achieve the same value for the odor unit: 1 $ou_E/m^3 \equiv 40$ ppb/v n-butanol.

These steps were taken in the full knowledge that the notion of a panel being representative of the general population was abandoned. The working group recognized that to be statistically representative, a panel should be far greater than the panel size of 5–8 that is practical for routine measurement.

With the steps mentioned above, a structure for quality assessment and control (QA/QC) could be applied based on the international standard ISO5725. This helps laboratories to maintain "calibration" and achieve a defined level of quality, based on assessor performance on the reference odorant. It has since been confirmed that the result of these steps to reduce variability in results and convergence between laboratories is transferable to other, environmental odors. This means that panel selection and quality control using a reference odor is an effective tool for anchoring the results of olfactometry and is indispensable for the QA/QC of the results for all odors obtained in a laboratory. The full benefit of the application of the quality criteria will become more apparent after long term application of the protocol, as described in the CEN draft standard. When the following protocol is applied for a sufficient period of time, the performance criteria as set in the CEN draft are attainable:

- Accuracy, $A_{od} \leq 0.217$ (calculated from the logarithms of test results): In 95% of the cases, results shall lie within less than a factor of 1.6 away from the assigned value of 40 ppb/v butanol.
- Repeatability, $r \leq 0.477$ (calculated from the logarithms of test results): In 95% of the cases, two single measurement results shall be less than a factor of 3 apart.

This defined quality of measurement results is compatible with the use of olfactometry in odor policy and regulatory applications.

REFERENCES

1. Prokop, W.H. "Developing odor control regulations: Guidelines and considerations," *J. Air Pollution Control Assoc*. **1977**, *28:1*, 9-16
2. Winneke, G.; Kastka, J. Odor Pollution and Odor Annoyance Reactions in the Industrial Areas of the Rhine-Ruhr Region. In *Proceedings of the 6th International Symposium on Olfactometry and Taste*; Information Retrieval Ltd.: London, 1977.
3. Lindvall, T. "On sensory evaluation of odorous air pollutant intensities," *Nordisk Hygienisk Tidskrift-Supp. 2*: **1970** Stockholm.
4. Berglund, B. "Quantitative approaches in environmental studies," *Int. J. Psychol*. **1977**, *12:2*, 111-113.
5. Berglund, B. "On the scaling of annoyance due to environmental factors," *Environ. Psychol. and Nonverbal Behaviour* **1977**, *2:2*, 83-92.
6. Winneke, G.; Kotalik, J.; Heldenich, H.O.; Kastka, J. "Zur wahrnehmung von schwefelwasserstofff unter labor und feldbedingungen," *Staub Reinh. der Luft* **1979**, *39*,156-159.
7. Berglund, B.; Berglund, U.; Ekman, G.; Engen, T. "Individual psychophysical function for 28 odorants," *Perception and Psychophysics* **1971**, *9*, 379-384.
8. Finney, D.J. *Probit Analysis*; Cambridge University Press: 1971.
9. Drake, D. "A Fortran program 'Simplot' for fitting sigmoid curves to threshold data for plotting results," *Chemical Senses and Flavor* **1975**, *1*, 519-531.
10. Dravnieks, A.; Schmidtsdorff, W.; Meilgaard, W. "Odor thresholds by forced-choice dynamic triangle olfactometry: Reproducibility and methods of calculation," *J. Air Poll. Control Assoc*. **1986**, *36*, 900-905.
11. Mills, J.L.; Walsh, R.T.; Luedtke, K.D.; Lewis, K.S. "Quantitative odor measurement." *J. Air Poll. Control Assoc*. **1963**, *13*, 467-475
12. Dravnieks, A.; Jarke, F.H. "Odor measurement by dynamic olfactometry: Significant operational variables," *J. Air Poll. Control Ass*oc. **1980**, *30*, 1284–89.
13. Thiele, V.; Kastka, J.; Winneke, G.; Prinz, B. "Zum problem der geruchsmessung: vergleichende untersuchungen zwischen MIU und LIS," *Staub Reinh. der Luft* **1981**, *41*, 289-295.
14. Köster, E.P. Limitations Imposed on Olfactometric Measurements by the Human Factor. In *Odor Prevention and Control of Organic Sludge and Livestock Farming*; Nielsen,V.C., Voorburg, J.H.,L'Hermité, P., Eds; Elsevier Applied Science Publishers: London and New York, 1985.
15. Hangartner, M.; Hartung, J.; Paduch, M.; Pain, B.F.; Voorburg, J.H. "Improved recommendations on olfactometric measurements," *Environ. Technol. Letters* **1989**, *10*, 231-236.
16. Ruth, J.H. "Odor thresholds and irritation levels of several chemical substances: A review," *Amer. Indus. Hygiene Assoc*. **1986**, *47*, 142-151.
17. van Harreveld, A.P. Introduction of Odor Regulations in the Netherlands and Related Practical Aspects. In *Proceedings of the International Specialty Conference on Recent Developments and Current Practices in Odor Regulations, Controls and Technology*; Air & Waste Management Association: Pittsburgh, PA, 1991; pp 34-38,.
18. Lindvall, T. "Methods for measuring and evaluating odorous air pollutants at the source and in the ambient air," *Nordisk Hygienisk Tidskrift* **1970**, *51:2*, 1-69.
19. *Industrial Odour Control: Guidelines for Environmental Authorities Administration of Odour Control Measurements in Industry, No. 4*; Ministry of the Environment: Copenhagen, Denmark, 1985.
20. Moss, R.L. *Development in the Measurement and Control of Process Odors. 50th Annual Conference*; National Society for Clean Air: Torquai, UK, 1983.
21. *VDI 3881 Part 1-4, Richtlinien,* (draft) *Olfactometry, Odour Threshold Determination, Fundamentals*; Verein Deutsche Ingenieure Verlag: Düsseldorf, Germany, 1980.
22. Roos, C.; Joziasse, J.; Klarenbeek, J.; Punter, P.H.; Schaefer, J. *Standaardisatie Olfactometers*. Publikatiereeks Lucht, No. 49; Ministry of the Environment: Leidschendam, the Netherlands, 1985.

23. AFNOR, (draft) *Method of the measurement of the odor of a gaseous effluent, X-43-101*; Bureau de Normalisation: Paris, 1981.
24. Thiele, V. "Experimentelle untersuchung zum ermittlung eines geruchsschwellenwertes für schwefelwasserstoff," *Staub Reinh. der Luft* **1979**, *39*, 159-160.
25. Thiele, V. "Olfaktometrie von schwefelwasserstoff: ergebnisse des VDI-Ringversuches," Staub Reinh. der Luft **1982**, *42*, 11-15.
26. Bahnmüller, H. "Olfaktometrie von dibutylamin, acrylsaüremethylesther, isoamylalkohol und eines spritzverdünners für autolacke - ergebnisse eines VDI-Ringvergleiches. *Staub Reinh. der Luft* **1984** *44*, 352-358.
27. Thiele, V. "Olfaktometrie an einen emissionsquelle: Ergebnisse des VDI – Ringvergleiche," *Staub Reinh. der Luft* **1984**, *44*, 342-351.
28. Thiele, V.; Bahnmüller, H. "Prüfung und standardisierung von olfaktometrische messungen," *Staub Reinh. der Luft* **1985**, *45*, 200-203.
29. Dollnick, H.W.O.; Thiele, V.; Drawert, F. "Olfaktometrie von schwefelwasserstoff, n-butanol, isoamylalkohol, propionsaüre und dibutylamin-ergebnisse eines VDI-Ringvergleiches. *Staub Reinh. der Luft* **1988**, *48*, 325-331.
30. Both, R.; Geueke, K.J.; Essers, K.H.; Otterbeck, K. *Ergebnisse eines Olfaktometrie-Ringversuches*. Aus der Tätigkeit der LIS Essen, Germany, 1992.
31. Hermans, L. *Ringonderzoek Olfactometers;* Publikatiereeks Lucht No.80; Ministry of the Environment: Leidschendam, the Netherlands, 1989.
32. Heeres, P.; Harssema, H. "Progress of the standardization of olfactometers in the Netherlands," *Staub Reinh. der Luft* **1990**, *50*, 185-187.
33. Harssema, H.; Heeres, P. A Long-duration Ring Test of Olfactometers. In *Biotechniques for Air Pollution Abatement and Odour Control Policies*; Dragt, A.J., Ilam, J., Eds.; Elsevier Applied Science: London and New York, 1992.
34. Heeres, P.; Harssema, H. *An Interlaboratory Study on the Odor Measurements of n-Butanol. Odours-Control, Measurements and Regulations;* Seminar Odours; Ekochem Ltd.: Szczecin, Poland, 1993.
35. Heeres, P.; Harssema, H. "An interlaboratory comparison of odor measurements of n-Butanol," *Gefahrstoffe Reinh. der Luft* **1996**, *56*, 55-60.
36. *Sensory Odour Measurements Using an Olfactometer - The Measurement of Odour Threshold Values*; Netherlands Normalization Institute: Delft, The Netherlands, 1990.
37. Harreveld, A.P.; Heeres, P. "The validation of the draft European CEN standard for dynamic olfactometry by an interlaboratory comparison on n-butanol," *Gefahrstoffe Reinh. der Luft* **1997**, *57*, 393-398.
38. Hemeon, W.L.C. "Technique and apparatus for quantitative measurement of odor emissions," *J. Air Poll. Control Ass*oc. **1968**, *18*, 166.
39. Dravnieks, A. Instrumental Aspects of Olfactometry Methods. In *Olfactory Research*; Moulton, Turk, Johnson, Eds.; Academic Press: New York, 1975.
40. American Society for Testing and Materials. *Standard Test Method for Measurement of Odor in Atmospheres (Dilution Method)*; ASTM: Philadelphia, PA, 1978; ASTM D-1391.
41. American Society for Testing and Materials. *Determination of Odor and Taste Thresholds by a Forced-Choice Ascending Concentration Series Method of Limits;* ASTM: Philadelphia, PA, 1979; ASTM Standard Practice E679-91.
42. Technical Research Centre of Finland, P.O. Box 204, SF-02151 Espoo, Finland.
43. University of Stockholm, Department of Psychology, S-106 91, Stockholm, Sweden.
44. DK-Teknik, Gladsaxe Möllevej 15. DK Technical 2860 Søborg, Denmark.
45. Silsoe Research Institute, Silsoe,Bedford MK45 4HS, UK.
46. Project Research, Singel 97, NL-1012 VG, Amsterdam, the Netherlands.
47. TNO, P.O. Box 342, NL-7300 AH Apeldoorn, the Netherlands.
48. LIS, Wallneyer str. 6, D-45133 Essen, Germany.
49. Mannebeck, Eckernfoerder str. 226a, D-2300, Kronshagen, Germany.
50. University Gent, FLTBW, Coupure Links 653, B-9000 Gent, Belgium.
51. CEA-CEN Fontenay, B.P. 6, F-92265 Fontenay-aux-Roses, France.
52. Institut für Hygiene, Klaussiusstr. 21, ETH-Zentrum, CH-8092 Zürich, Switzerland.
53. Dravnieks, A.; Prokop, W.H. "Source emission odor measurement by a dynamic forced choice triangle olfactometer," *J. Air Poll. Control Ass*oc. **1975**, *25*, 28-35.
54. Schaefer, J. *Vergelijking van Olfactometers voor Stankonderzoek*; CIVO-TNO Report No. A81.170/170052, 1981.
55. Laing, D.G. "Characterization of human behaviour during odour perception," *Perception* **1982**, 11, 221-230.
56. Feenstra, L. *Onderzoek verdunning ruikbekers;*. Intern report MT-TNO. No. 80-04274 TNO: Apeldoorn, the Netherlands, 1980.
57. Dupraz, C.A. "Observations on the impact of flow rates on odor thresholds," *Int. Soc. for Respir. Protection J.* **1990**, *1*, 34-36.
58. O'Brian, M.; Duffee, R.A.; Ostojic, N. Effect of Sample Flow Rate in the Determination of Odor Thresholds. In *Proceedings of Odors: Indoor and Environmental Air*; Air & Waste Management Association: Pittsburgh, PA, 1996.
59. Punter, P.H. "Measurement of human olfactory thresholds for several groups of structurally related compounds," *Chemical Senses* **1983**, *7*:.3/4, 215-235.
60. Wilby, F.V. "Variation in recognition odor threshold of a panel." *J. Air Poll. Control Ass*oc. **1969**, *19*, 96-100.
61. *ISO 5725: Precision of test methods - determination of repeatability and reproducibility for a standard test method by inter-laboratory tests, Second Edition;* International Standards Organization, 1986.
62. *CEN/TC264/WG2 Draft European CEN Standard: Odour Concentration Measurement by Dynamic Olfactometer*, CEN: Delft, the Netherlands, Revised 1995.
63. Harreveld, A.P.; Heeres, P. "Quality control and optimization of dynamic olfactometry using n-Butanol as a standard reference odorant," *Staub Reinh. der Luft* **1995**, *55*, 45-50.
64. Laska, M.; Hudson, R. "A comparison of detection thresholds of odour mixtures and their components," *Chemical Senses* **1991**, *16*, 651-662,

About The Authors

Ton van Harreveld is founder and director of Project Research Amsterdam BV (the Netherlands) and Project Research Environmental Consultants Ltd. (UK), both consulting firms specializing in odor characterization and odor annoyance management. He is acting convenor of the Working Group 2 "Odours" of CEN/TC264 Air Quality of the Committee Européen de Normalisation (CEN). Paul Heeres and Hendrik Harssema are senior researchers and lecturers at the Wageningen Agricultural University, Department of Air Quality. Two of their long-standing research subjects are odor measurement and atmospheric dispersion modeling.