CIÊNCIA E TÉCNICA VITIVINÍCOLA
Ciência Téc. Vitiv. 40(1) 10-28. 2025

Article

TRIANGLE TEST IN SENSORY ANALYSIS: APPROACHES AND CHALLENGES IN THE CONTEXT OF PROFICIENCY TESTING

TESTE TRIANGULAR EM ANÁLISE SENSORIAL: ABORDAGENS E DESAFIOS NO CONTEXTO DE ENSAIO DE APTIDÃO

Manuel Pinto[1]*, Paulo Barros[1], Elisete Correia[2], Alice Vilela[3]

[1]ALABE – Associação dos Laboratórios de Enologia, Rua de Ferreira Borges, 27, 4050-253 Porto. Portugal.

[2]Chemistry Research Center (CQ-VR), Dep. of Agronomy, School of Agrarian and Veterinary Sciences, University of Trás-os-Montes and Alto Douro, Vila Real, Portugal.

[3]Center for Computational and Stochastic Mathematics (CEMAT), Dep. of Mathematics, University of Trás-os-Montes and Alto Douro, Apt. 1013, 5001-801, Vila Real, Portugal.

***** Corresponding author: Tel: +351 935576384; email: mmnpinto@gmail.com

(Received 23.09.2024. Accepted 13.02.2025)

SUMMARY

Participation in proficiency testing (PT) programs, as outlined in ISO/IEC 17025, is a vital tool for ensuring the validity of laboratory results. Although it requires an initial investment, the benefits—such as reduced errors, improved efficiency, and the prevention of costly problems—make it a cost-effective approach. This participation enhances accuracy, saves costs, and increases laboratory productivity. The SENSORIAL-ALABE test is designed to enhance tasters' sensory abilities, offering sensory panels or individual assessors the unique opportunity to track their performance over time confidentially. For this purpose, the triangle test is used, in which three samples are presented to the evaluator in different orders, two of which are identical. The evaluator's task is to identify the different sample, a process crucial in testing the sensory acuity of the assessor and the group. This method is essential for quantitatively evaluating the response to progressively increasing olfactory stimuli related to defects or aromas in wines and/or wine spirits. The assessor performs the test in four increasing concentrations, identifying the different samples and the compound used based on a table of compounds and sensory descriptors previously established. The results reflect the overall performance of the assessor in the four triangle tests at progressively increasing concentrations, using binomial distribution to assess statistical significance. This study examined the effects of adding a compound to one or two samples to examine whether it influenced participants' accuracy in identifying the correct samples. Results suggest that adding the compound to two samples generally makes it more challenging for participants to accurately identify the odd sample, leading to more incorrect answers. The proficiency tests show a continuous improvement in the tasters' performance, especially when they face the same challenge a second time. These tests are essential for constantly improving laboratories performance, enhancing assessors' sensitivity through training, and providing relevant information for their qualification.

RESUMO

A participação em programas de testes de aptidão, conforme delineado na ISO/IEC 17025, é uma ferramenta fundamental para garantir a validade dos resultados laboratoriais. Embora exija um investimento inicial, os benefícios - como a redução de erros, melhoria da eficiência e prevenção de problemas dispendiosos—tornam esta abordagem rentável. Esta participação não só melhora a precisão, como também leva a poupanças e ao aumento da produtividade do laboratório. O SENSORIAL-ALABE é um teste destinado a melhorar as capacidades sensoriais dos provadores, permitindo que painéis sensoriais ou provadores individuais monitorizem o seu desempenho ao longo do tempo de forma confidencial. Para tal, é utilizado o teste triangular, no qual três amostras são apresentadas ao avaliador em diferentes ordens, sendo duas delas idênticas. O avaliador deve identificar a amostra diferente. O objetivo é testar a acuidade sensorial do provador e do grupo. O teste avalia quantitativamente a resposta a estímulos olfativos progressivamente crescentes, relacionados com defeitos ou aromas em vinhos ou bebidas espirituosas. O provador realiza o teste em quatro concentrações crescentes, identificando a amostra diferente e o composto utilizado, com base numa tabela de compostos e descritores sensoriais previamente estabelecida. Os resultados refletem o desempenho global do provador nos quatro testes triangulares em concentrações progressivamente crescentes, utilizando a distribuição binomial para avaliar a significância estatística. Este estudo examinou os efeitos da adição de um composto a uma ou duas amostras para verificar se influenciava a precisão dos participantes em identificar as amostras corretas. Os resultados sugerem que adicionar o composto a duas amostras torna geralmente mais difícil para os participantes identificar corretamente a amostra diferente, originando uma taxa de respostas erradas mais elevada. Os testes de aptidão revelam uma melhoria contínua no desempenho dos provadores, especialmente quando enfrentam o mesmo desafio uma segunda vez. Estes testes são essenciais para a melhoria contínua da performance dos laboratórios, aumentando a sensibilidade dos provadores através de treino e fornecendo informações relevantes para a sua qualificação.

Keywords: Triangle test; assessors; discrimination test; quality control; proficiency tests.
Palavras-chave: Teste triangular; provadores; teste de discriminação; controlo da qualidade; ensaios de aptidão.

Article available at https://www.ctv-jve-journal.org or https://doi.org/10.1051/ctv/ctv2025400110

INTRODUCTION

Participation in proficiency testing (PT) programs is one of the available tools for assuring the validity of results that laboratories can use, as laid down in ISO/IEC 17025 (ISO 17025, 2017). It is also one of the most efficient and cost-effective mechanisms to achieve this goal since, although there may be an initial investment associated with participation, the benefits, such as reduced errors, improved efficiency, and prevention of more costly problems, significantly reduce the risk of errors in the execution of the activity that can lead to incorrect results and thus wrong decisions. Therefore, participation in proficiency testing also contributes to cost savings and improved laboratory productivity.

However, for a laboratory to derive maximum benefit from its participation, whatever the purpose, it must ensure that it selects proficiency testing programs that give it confidence in the results and suit its own needs. Selection of a proficiency supplier, intercomparison or proficiency testing selection and evaluation of proficiency testing results are three key aspects that every laboratory should consider when designing its participation in proficiency testing programs (Kilcast, 2010). The competence of proficiency test suppliers is critical, as poorly managed tests with inadequate data processing or unstable samples can lead laboratories to incorrect conclusions. Laboratories must verify the technical competence of suppliers to ensure tests meet their needs. ISO/IEC 17043 (ISO, 2023) specifies the requirements for proficiency testing providers, and accreditation under this standard demonstrates competence in designing, planning, and conducting proficiency tests within the accredited scope. Although the Portuguese Accreditation Institute (IPAC) does not accredit providers under ISO 17043, laboratories must ensure supplier competence, following guidelines in IPAC's DRC005 document (IPAC, 2019). Laboratories participate in intercomparisons for various objectives, such as evaluating performance, identifying issues with methods or equipment, assessing personnel training and supervision, and ensuring comparability of results. To achieve these goals, laboratories must select proficiency testing providers that meet their specific needs. Before registering, it is crucial to analyze all relevant information about the intended proficiency test. Clearly defining the objective of participation helps laboratories identify their requirements and select the most suitable test. Laboratories should evaluate the results of their participation in proficiency testing for both satisfactory and unsatisfactory results because it will allow them to derive maximum benefit by establishing actions aimed at continuous improvement through the identification of corrections or adjustments to their measurement processes. The key to practical proficiency testing results evaluation is to interpret all the information provided by the supplier in the results reports they issue and not focus only on the supplier's performance evaluation (e.g., z-score value or compatibility indices). Proficiency testing reports contain relevant technical information that enables laboratories to execute this task.

Laboratories should also confirm that proficiency testing has met the expectations and needs in their initial planning. Suppose a proper evaluation of participants' performance in a proficiency test is not conducted; in this case, the test may fail to provide valuable insights or even worse, create a false sense of security for the laboratory, potentially leading to significant issues.

The better laboratories understand proficiency testing, the better they can benefit from it, ensuring it is useful and its results are technically reliable, while avoiding false security.

Proficiency testing is critical for sensory panels because it ensures reliable and consistent performance beyond individual assessor validation. While validating assessors is necessary, it does not guarantee overall panel reliability. Therefore, sensory panels must participate in proficiency tests to ensure their collective output meets the required standards. Participation in a PT scheme is the most effective way to demonstrate the ability of a sensory panel and is essential for accredited laboratories. PT performance is critical for showcasing the repeatability and reproducibility of sensory results, which form the basis for decisions about product quality. While individual assessor performance has been studied extensively, the performance of the panel as a whole has deserved less attention. Nevertheless, the panel's overall performance depends on the contributions of its members, as poor performance by assessors negatively impacts the panel's effectiveness (McEwan *et al.*, 2003; Hyldig, 2010).

Sensory evaluation tests can be categorized into two basic groups: analytical tests (trained panels) and affective tests (consumers) (Duizer and Walker, 2016). Difference tests, classified under analytical tests, determine whether a sample can be distinguished at a given significance level (typically 5%) (Kemp *et al.*, 2009). The most used difference test models are paired comparison, triangle, duo-trio, ranking, and multiple comparisons (Kemp *et al.*, 2009; Stone *et al.*, 2012; Palermo, 2015).

Discriminative tests can be used to achieve several of practical objectives. In some cases, the interest lies in demonstrating that two samples are perceptibly different. In other cases, the interest lies in determining whether two samples are similar enough to be used interchangeably (Meilgaard *et al.*, 2015).

Discriminative tests also assess individuals' ability to discriminate various stimuli for selection in trained analytical panels (for example, descriptive or quality control panels) (Ojeh, 2020). All discriminative tests aim to determine whether an assessor can detect a difference between the products being analyzed through the stimuli captured by their senses. If the products generate doubts and a difference between them can be detected, then there is no point in carrying out the test (Ford, 2017; Sinkinson, 2017). As stated by Kemp *et al.* (2024), differences between products should be slight: there is no point in conducting a discriminative test on products that are different. In decision-making based on experimental evidence, errors are inevitable: rejecting the null hypothesis when it is true leads to a type I error, while failing to reject it when it is false results in a type II error. The difference test determines if there is a noticeable difference between two samples, focusing primarily on minimizing the α-risk (type I error). The number of evaluators is chosen based on the α-risk table and practical considerations, such as sample and evaluator availability. However, the β-risk (type II error) and the proportion of distinguishers (pd) are often ignored or considered unimportant, leading researchers to accept higher β-risk values to keep the panel size manageable (Meilgaard *et al.*, 2015).

For many types of sensory methods, references prepared in the laboratory from chemical compounds (for example, molecules that elicit sensory responses) of known purity and composition can be used; in other methods, samples to which substances are added may be necessary to produce specific qualities, defects, or positive attributes. The availability of Certified Reference Materials (CRM) is minimal (there is a limited supply or access to CRMs for the intended use); on the other hand, samples from interlaboratory tests can be used when they exist, although the laboratory must assign a value based on the available information. This type of material is very suitable for quality control of sensory methods (EA, 2022).

Few structured proficiency tests for sensory discriminative tests aroma recognition are widely available, which limits the ability to train and evaluate sensory panelists effectively. The absence of sensory PT schemes can be ascribed to the challenges involved in establishing criteria for evaluating panel performance (McEwan *et al.*, 2003).

This project focused on designing and implementing a proficiency test to assess assessors' discrimination ability in sensory analysis. It aimed to report on the experience gained in organizing and conducting proficiency tests, emphasizing performance evaluation to ensure accuracy and reliability. The study critically analyzed the participation of assessors from different laboratories of sensory analysis in the SENSORIAL-ALABE test, organized by Associação dos Laboratórios de Enologia (ALABE). The test assessed assessors' responses to olfactory stimuli, helping in training and standardization. It also examined trends in assessor performance over a decade, highlighting challenges in olfactory recognition and the impact of training. The findings contribute to improving training methods and strengthening sensory analysis.

MATERIALS AND METHODS

The test has been running since 2009, and this study will consider data collected up to 2022. Three testing sessions, referred to as editions, were conducted each year. Each edition included three compounds until 2017, and only two from 2018 onwards. The assay used in the SENSORIAL-ALABE is a triangle test designed to assess discrimination ability. It identifies a difference between samples and measures the individual's ability to detect sensory characteristics.

This proficiency test is designed to quantitatively assess the response to olfactory stimuli of progressively increasing intensity. These stimuli are typically associated with detectable defects or aromas in wines or wine spirits, but they can be found in various products. The test serves as a tool for maintaining, improving, or training the sensory abilities of assessors. It allows individual and collective performance to be monitored under a strict confidentiality regime. Results are processed anonymously, with only the group coordinator and/or the assessor accessing their performance. By conducting these evaluation tests, sensory panels or individual assessors can better understand their performance and track its evolution.

The proposed scheme is similar to the interlaboratory tests usually carried out in laboratories. The results are sent to ALABE, which processes them under a strict anonymization scheme. ALABE publishes a report in which only each person can be identified by the code they have defined for themselves. Only participants/groups of assessors duly registered with ALABE may participate in this test. Individual assessors or groups must carry out the tests by the specific procedure adopted.

Planning of distribution and preparation of samples

According to the previously established schedule, ALABE distributes sets of vials for each edition containing two compounds (from 2009 to 2017, three were used in each edition). All the participating organizations received the same samples. It is assumed that the test will take place under the same conditions (same compounds, same order, e.g., two positives and one negative; same glasses, among others) for sensory evaluation, thus being on equal

assessment so that a final and comparable judgment can be made.

Tasting panel

Between 2009 and 2022, 54 entities and 831 assessors were involved, and 26 compounds were analyzed. The entities that took part in this test were directly or indirectly linked to the wine sector, and their assessors were members of accredited or non-accredited tasting panels.

Assay methodology

The triangle test determines whether there is a perceptible sensory difference between two homogeneous products in one or more sensory attributes. It does not identify or determine the difference's nature, size, or direction. Triangle testing is a widely used and easily understood method for assessors. It is applicable in most scenarios, although its effectiveness is limited when dealing with products that cause fatigue, carryover effects, or adaptation, as well as with assessors who may find evaluating three samples confusing. In the triangle test, three samples are presented simultaneously or sequentially; two are identical, and one is different. The evaluator is instructed to taste in the order given and choose the "odd" or "different" sample; therefore, the question asked is: "Which sample is different?" With three samples presented, the task seems simple but is often confusing, mainly when slight differences exist. The probability of giving a correct answer (p) is one-third, while the likelihood of giving an incorrect answer (q) is two-thirds. This test is one-tailed compared to the equal-different test, with p and q= 1/2, which is a two-tailed test. The triangle test is considered less sensitive in rejecting H_0. It requires more evaluators for a given difference compared to the equal-different or tetrad tests and, therefore, has higher cost expenses. It is, however, statistically more efficient than the duo-trio test. The analysis involves counting the number of correct identifications of the odd sample and the total number of responses. Responses indicating "no difference" are not counted as valid. If the odd sample is not detectable, assessors should be encouraged to guess, as the analysis accounts for the possibility of correct responses occurring by chance (Civille *et al.*, 2024).

This proficiency test trial aims to evaluate the sensory acuity of the assessor, their group, and the overall participating population. In this test, the samples comprised aqueous solutions of high-purity compounds in different concentrations. The methodology used was based on multiple triangle tests. According to ISO 4120 (ISO 4120, 2021), the triangle test determines whether there are differences between two samples when the products provoke simple, weak stimuli. The test is designed to quantitatively assess the response given by assessors to sensory stimuli of progressively increasing intensity. The assessor is instructed to conduct a triangle test using olfactory sensory analysis at four progressively increasing concentrations, with the understanding that each set consists of two identical samples and one different sample. The (coded) samples must be evaluated in the order provided, and the assessor is tasked with identifying the different sample in each set. A forced-choice approach is applied in cases where the assessor is unable to determine the different sample. At the same time, the assessor is asked to identify the compound (since 2018) used based on a table of possible compounds and their sensory descriptors, as shown in Table I.

To summarize, ALABE delivers vials to be distributed among the tasting glasses, ready to dilute their contents in an appropriate volume of water. The assessor is then asked to carry out a triangle test using only olfactory sensory analysis at four concentrations and intensity pretended perceptions, e.g.: (1) low concentration, close or equal to the theoretical olfactory perception threshold; (2) second level of concentration; (3) third level of concentration; (4) fourth level of concentration, generally perceived by most individuals. At each concentration level, one of the three glasses contains a solution that differs from the other two, with the presentation order randomized (e.g., "positive-positive-negative" or "positive-negative-negative"). The distribution of positive and negative samples varies across editions. Standardizing sensory Quality Control (QC) procedures is crucial to ensure consistency and minimize procedural variability so that any observed differences can be ascribed to the samples. Within a company, standardization should address sample handling (e.g., storage, preparation, and service), panel management (e.g., number of assessors, evaluation timing, and panel reliability checks), and data analysis and communication. Procedures for data analysis and communication include standardizing result interpretation and sharing findings with relevant personnel. Communication methods differ based on whether a sample meets the required criteria or does not. Results that meet the required criteria may be included in a weekly summary, while samples that do not meet the criteria require immediate action and the involvement of additional stakeholders to determine product disposition. Clear, agreed-upon specifications are essential to streamline decision-making and reduce disputes (Schultz, 2021).

The testing methodology follows the protocol (ALABE, 2024) adopted and distributed to all participants, explained below.

Timetable

The assay has three editions per year, each with two compounds available, initially with three compounds. The reduction to two compounds per

edition was mainly due to the logistics of the material needed to carry them out and various reasons relating to the schedules of multiple organizations, thus making it easier to meet the deadlines set to performing the test. For laboratories that have been accredited for a specific analytical method, the accreditation bodies demand that the laboratories participate in proficiency testing at regular intervals (Kilcast, 2010). Concerning the minimum frequency of participation by laboratories, it should not be less than one representative participation (e.g., by type of product, characteristic, and technique) of the accredited scope during each accreditation cycle (4 years) (IPAC, 2019).

Table I

Possible compounds and their sensory descriptors when in aqueous solution

Compound	CAS ID	Purity (%)	Aroma descriptor	Detection Sensory Threshold (ppb)
Pyridine	110-86-1	99.8	Fishy (Yeretzian, 2017)	77 (Yeretzian, 2017)
1-Hexanol	111-27-3	97	Cut grass, green, dried leaves, ethereal fruity green, sweet, alcoholic linseed oil (Grainger, 2021)	200 to 2500 (Burdock, 2016)
2,4,6-Trichloroanisol	87-40-1	99.5	Mold, damp, wet cardboard (Grainger, 2021; AWRI, 2024)	0.00003 - 0.0003 (Curtis et al., 1972; Griffiths, 1974; Malleret et al., 2001)
2,6 – Lutidine	108-48-5	98	Minty-tarry odor (Burdock, 2016)	n/a
2-Mercaptoethanol	60-24-2	99+	Reduction aromas (gas, rotten eggs) (AWRI, 2024)	120–640 (Greim, 2024)
2-Methylisoborneol	2371-42-8	99.9	Mold; dirty, earthy, moldy water (Grainger, 2021; AWRI, 2024)	0.042 (Persson, 1980)
2-Octanol	123-96-6	97	Fresh spicy green woody herbal earthy (Burdock, 2016)	7.8 to 42 (Burdock, 2016)
2-Phenylethanol	60-12-8	98	Rose-like (Burdock, 2016)	0.015 to 3500; Recognition: 1200 (Burdock, 2016)
4-Ethylphenol	123-07-9	97+	Horse sweat, leather, band-aid, medicinal or pharmaceutical (Grainger, 2021; AWRI, 2024)	42 to 130 (Burdock, 2016)
4-Vinylphenol	2628-17-3		Pharmaceutical, pharmaceutical ink, gouache (Grainger, 2021)	10 to 85 (Burdock, 2016)
Acetaldehyde	75-07-0	99	Pungent, ethereal, fresh, lifting, penetrating, fruity, musty (Yeretzian, 2017)	0.7 (Yeretzian, 2017)
Benzaldehyde	100-52-7	99+	Bitter almond (Grainger, 2021)	100 ppb to 4600 (Burdock, 2016)
Citral	5392-40-5	95	Strong, lemon-like odor, bittersweet taste (Burdock, 2016)	0.01 (Burdock, 2016)
Coumarin	91-64-5	99.9	Sweet, fresh, hay-like, odor similar to vanilla seeds, burning taste with bitter undertone, nutlike flavor on dilution (Burdock, 2016)	34 to 50; Recognition, 250 (Burdock, 2016)
Diacetyl	431-03-8	97	Buttery, creamy, fatty, oily, sweet, vanilla (Yeretzian, 2017)	0.3 (Yeretzian, 2017) 0.3 to 15; Recognition: 5 (Burdock, 2016)
Ethyl acetate	141-78-6	99.8+	Glue, nail varnish/acetone (Grainger, 2021; AWRI, 2024)	5 to 5000 (Burdock, 2016)
Eucalyptol	470-82-6	99	Weet, cooling, fresh, chemical pine, slightly minty with a spicy cardamom nuance (Burdock, 2016)	1 to 64 (Burdock, 2016)
Eugenol	97-53-0	99	Clove, spice (Grainger, 2021)	6-30 (Diep *et al.*, 2021)
Fenchone	4695-62-9	99.5+	Cooling camphoreous, terpy, spicy, sweet mentholic pine-like (Burdock, 2016)	510 (Burdock, 2016)
Guaiacol	90-05-1	99.4	Phenolic, burnt, smoke, spice, vanilla, woody (Yeretzian, 2017)	2.5 (Yeretzian, 2017) 3-31 (Burdock, 2016)
Isoamyl acetate	123-92-2	97+	Fruity, banana, sweet, fragrant (Burdock, 2016)	2 to 43 (Burdock, 2016)
Limonene	138-86-3	97	Citrus, herbal, terpene, camphor (Yeretzian, 2017)	4 (Yeretzian, 2017) 4 to 229 (Burdock, 2016)
Linalool	78-70-6	95+	Flowery, citrus, orange, terpene, waxy, rose (Yeretzian, 2017)	0.17 (Yeretzian, 2017) 4 to 10 (Burdock, 2016)
Vanillin	121-33-5	99	Sweet, vanilla, creamy (Yeretzian, 2017)	25 (Yeretzian, 2017) 29 to 1600; Recognition: 4000 (Burdock, 2016)

The homogeneity and stability of the proficiency test item are assessed beforehand. It consists of preparing a package with material samples and carrying out a trial shipment, ensuring that the distance and transportation are realistic. When the samples arrive, it is possible to check that the sensory quality is the same as when they were sent.

The aromas are distributed according to a pre-established timetable for each year, which contains the distribution dates, the deadlines for receiving the results, and the expected date for finalizing the report.

Although the ideal situation would be for all participants to carry out the test simultaneously, this is impractical given the many assessors and their very different locations. To minimize any phenomenon that could compromise the integrity of the tested chemical compound, the timeframe for carrying out the proficiency should be reduced. Likewise, other factors that could lead to deviations (dilution, thawing, heating, among others) should be considered, and sample preparation should be protocolized appropriately to ensure that the sensory characteristics remain identical from laboratory to laboratory.

Compound coding

The samples must be representative of the product and prepared similarly. In this test, the aim was for the sequences to be randomized and for there to be a balance between "positive" samples (with added compound) and "negative" samples (water) (Table II). ALABE delivers sets of vials for each edition containing two compounds commonly identified with detectable defects/aromas in wines and/or wine spirits (e.g., Compound A - 1st edition, 2023, Compound B - 1st edition, 2023, etc.). Vials are coded according to Table III.

Table II

Example of the samples' randomization in the four sequences of the triangle test

Sequences	Sample 1	Sample 2	Sample 3
Sequence 1	+	-	+
Sequence 2	-	-	+
Sequence 3	-	+	+
Sequence 4	+	+	-

Table III

SENSORIAL-ALABE vial coding

Sequences	Codes		
Sequence 1	A 1.1.	A 1.2.	A 1.3.
Sequence 2	A.2.1.	A.2.2.	A.2.3.
Sequence 3	A.3.1.	A.3.2.	A.3.3.
Sequence 4	A 4.1.	A 4.2.	A 4.3.

An accredited laboratory, Centro Tecnológico da Cortiça (CTCOR), is responsible for preparing the solutions following good laboratory practice. The quantity in each vial is strictly measured so that, if the dilution procedure recommended here is followed, the triangle test can be carried out correctly at four different and increasing concentrations. As this is a quantitative test, dilution must be carried out rigorously, putting assessors at the same difficulty level and allowing for results to be compared.

Aroma compound stability

The organization of proficiency tests faces significant challenges, primarily in maintaining the integrity of the product, particularly aromatic compounds, over the test duration. As highlighted by Kilcast (2010), preliminary tests are essential to evaluate the product's behavior over time.

Certified Reference Materials (CRMs) remain scarce, but interlaboratory test samples can sometimes serve as substitutes. Physical reference standards, often called "gold standards," are critical for sensory methods like difference tests and must exhibit all key attributes without aberrations. However, maintaining the quality of these standards is challenging and requires careful consideration. The challenge with a physical reference standard lies in preserving its quality, making it crucial to carefully consider how it can remain the gold standard over time. Research and development (R&D) can develop a new product to serve as a benchmark for comparison, or the gold standard can be stored using methods that minimize aging, such as freezing. The preservation method will largely depend on the product type, and initial testing may be necessary to determine the best approach. Regardless of the storage method, the reference standard should be regularly assessed and replaced once it no longer meets acceptable criteria (Schultz, 2021).

Strict guidelines should be followed to mitigate aroma preservation issues: store vials at 5°C ± 2ºC unless otherwise specified, avoid freezing, and keep them in packaging until use. Adopting a tight test timetable further reduces preservation risks.

Dilutions

The option adopted in the trial was to distribute a concentrated solution to make it easier to send to the participants without compromising the stability of the chemical compound. The water for the final dilution was specified since sensory differences in water used for dilution, even subtle ones, could compromise the PT. As some authors stated (Mitterer-Daltoé *et al.*, 2012)) regarding studies conducted to measure sensory thresholds with several variables (nasal occlusion of the assessor,

volume of aqueous sample, type of water, and statistical methods of analysis), the quality of the water influences the sensitivity to primary flavors and the metallic sensation. They point out that preference should be given to water with low mineralization.

To accurately perform the assay dilution procedure, the vials must be organized according to their aroma, identified by the initial letter of their code (e.g., A, B, etc.), and refrigerated for at least one hour to reduce volatility losses upon opening. Following this, 12 glass containers should be prepared for each compound, preferably using 250 mL ground glass flasks with lids, although stoppered flasks or beakers are also acceptable. Each container should be filled to the 250 mL mark with commercially available water that is neutral in odor and contains less than 10 mg/L of calcium, ensuring the same batch of water is used in all dilutions. The containers must be labeled according to the table for compound "A," with the same labeling process applied to each subsequent compound. Using nitrile gloves, the vial within the numbered sachet should be uncapped and placed directly into the container along with the screw capsule, taking care to avoid contamination of the vial surfaces. It is essential to work methodically from the lowest to the highest concentration, verifying the marking on the sachet before proceeding. The cap from the vial must be removed swiftly, holding the vial over the container's neck and allowing it to drop upright into the container, with the screw capsule inserted as well. The vial should remain in the container, as its volume is factored into the concentration calculation. Once the vial is inserted, the container must be closed, and the dilution homogenized by shaking the solution continuously for at least 15 s. This procedure should be repeated for compound B and any other compounds. Upon completion of the dilutions, the test should be conducted immediately.

Test preparation

The test preparation involved several steps. First, a sufficient number of sets of 12 glasses were prepared to match the number of assessors participating in the test and tasting sequence. Each set of glasses was then labeled according to the identifiers provided in the corresponding lines of Table III, with three glasses marked for each line. Next, the solutions listed in the first line of Table III were evenly distributed among the glasses, with 40 mL per glass. Finally, the glasses were placed in the test booths, ensuring the solutions were kept at 20 ºC ± 3 ºC.

Test execution

The test was carried out in accordance with the guidelines issued in the ISO 4120 (ISO 4120, 2021). The assessors should first be directed to test the lowest concentration of the compound. This concentration corresponds to the vials listed in the first line of each table, where the central character is "1." For instance, "A 1.2." denotes compound A at the lowest concentration, precisely the second of three replicates. After the initial test, the glasses listed in the second line of Table III should be provided, with the previous set removed. This procedure is repeated until all glasses have been tested. The triangle tests are to be conducted in isolation, with discussions limited to those necessary for understanding the test, and any further comments should be reserved until all assessors have completed the tests. It is essential to use test glasses that conform to the standards set in ISO 3591 (ISO 3591, 1977).

Communication/submission of results

Each organization had a "participation code" (password) that was sent in good time to report results to ALABE. ALABE analyzed the individual and/or group data and sent a report to the assessor or group coordinator.

When entering data online, the assessor(s) was coded with an 8-character password of letters and numbers (e.g., A345Y97F). The first character was a letter and was kept in each edition. This was the code with which the assessor recognized himself in the ALABE report.

Each entity entered the data directly on ALABE's website. For this purpose, ALABE provided a "username" and "participation code" (password). ALABE has kept the codes and results submitted by the participants confidential.

Statistical analysis and treatment

A simplistic approach to demonstrate product differences involves applying the binomial test using all available observations, such as nk observations, when the same number of replications is obtained from each assessor (Meyners and Carr, 2024). This method is valid for identifying product differences (Kunert and Meyners, 1999). Consequently, if the binomial test shows a statistically significant result, it is reasonable to conclude that the products are significantly different. Essentially, under the null hypothesis of no product differences, it is impossible to differentiate between products, meaning assessor heterogeneity cannot occur (as all assessors share the same success probability of p_0). Given proper randomization and study execution, this ensures that all observations are independent, making the use of the binomial test appropriate.

In addition to calculating averages and the percentage of correct answers, the binomial test was used to validate the answers for each of the four triangle test sequences. The binomial test is the most used statistical method in triangle tests (O'mahony, 1986). It evaluates whether the number of correct

identifications (or "hits") is significantly greater than would be expected by chance. The null hypothesis assumes that the assessors are guessing, with a 1/3 probability of correctly identifying the odd sample. The binomial distribution calculates the probability of getting a given number of correct identifications.

Each of the four triangle test sequences was analyzed separately using the binomial test to identify statistical differences between the samples. However, the proficiency testing methodology focuses on the evaluator's overall performance across all four tests, emphasizing their ability and consistency rather than isolated test results. In this context, the sensory threshold of the assessor is defined as the lowest concentration at which they can consistently identify the different samples without errors in the higher concentration sequences. If the evaluator incorrectly identifies the different samples in the sequence with the highest concentration (sequence 4), their performance for that compound is recorded as 0% correct. Indeed, ALABE assumes that an error at the highest concentration indicates that correct identifications at lower concentrations were likely due to chance.

The analysis of the results of the triangle test (for each of the four concentration sequences used) is based on the binomial distribution, analyzing the number of correct answers and comparing them with the number needed to reach a specific statistical significance depending on the number of participants (Lawless and Heymann, 2010). The number of correct answers in each concentration sequence used was counted. Whenever this number is lower than the minimum (critical) number of correct answers according to the binomial criterion for a significance level of 5 %, it means that statistically, there were no differences between the samples in one of the concentration sequences used. Such situations were carefully analyzed by ALABE to understand the causes that could have led to this result. If the problem occurred only in the lowest concentration sequence (Seq.1/Conc.1), it could mean that the concentration used was below the olfactory detection threshold. However, if the situation occurred in a higher concentration sequence (Seq.2; Seq.3 or Seq.4) with Seq.1 being positively validated by the binomial criteria, this could constitute an anomalous situation caused by a ‘fault’ in the preparation of the compound which would invalidate the test.

Using high-performing taste assessors in various contexts can lead to a better understanding of consumer preferences, improved product quality, differentiation in the market, and fact-based decision-making (Beeren, 2018). The SENSORIAL-ALABE test's type highlights the importance of using top-performing assessors in sensory tests. Based on the type of the assay, as a discriminant test and regarding the importance/utility of using assessors with better performance in sensory tests in several contexts, the global ability to discriminate performance indicators was calculated using a specific formula (Equation 1).

Ability to discriminate = (Overall “gross” performance x Number of records)/Weighted Average **Eq. 1**

As described by Pinto *et al.* (2024), the formula adopted corresponds to a mechanism developed internally by the ALABE team to find a methodology that makes it possible to establish a ranking in terms of the performance of organizations and assessors, but at the same time considering the number of records (participations), and dividing by the weighted average, which consists of multiplying each value (average discrimination ability) by its weight (records). Finally, all the values are added together, and the result is divided by the sum of the weights (records). In essence, this method ranks assessors by balancing their individual performance and participation frequency, ensuring that those with more experience and consistently high scores are properly recognized.

The weighted average is calculated by proportionally combining each entity's performance with its number of records, giving an influence on entities with more records due to their more considerable representation. Overall “gross” performance corresponds to the sum of overall performance divided by the number of records (participations). For example, an assessor with 124 participations, a sum of overall performance equal to 11025, an overall “gross” performance equal to 88.9, and considering that the weighted average is 61.85, get a global ability to discriminate performance equal to 178.25. Ability to discriminate performance provides a comparative ranking of an assessor’s ability to discriminate, with higher values indicating better performance compared to others.

RESULTS AND DISCUSSION

The results of multiple triangle tests carried out over 13 years are presented in this work. The percentage of correct hits of the assessors that make up each entity was calculated (Table IV), and the average percentage of correct answers (% discrimination settlement) by entity is available (Table V).

In a triangle test, the probability of correctly identifying the sample among the three presented is $p = 1/3$. In the example shown in Table VI, corresponding to 217 assessors, statistical evidence indicates that the difference between the two samples is significant at a 5%. The analysis involves counting the number of correct answers. The difference between the samples was assessed using statistical

tables provided by Roessler *et al.* (1978), also referenced in other publications (ASTM E1885-04, 2011; Stone *et al.*, 2012; Meilgaard *et al.*, 2015; Rogers, 2017; ISO 4120, 2021), provide the minimum threshold of correct answers needed to confirm a significant difference between treatments.

Table IV

Example of results and percentage of correct answers in the four sequences of the triangle test for a given compound (geosmin)

Total number of entities: 23		KEY RESULTS				% CORRECT HITS
COD_ ENTITY	ID_ ASSESSOR	B.1.2.	B.2.3.	B.3.1	B.4.3.	
		RESULTS				
#3DEGT	GT567SHY 1	B.1.2.	B.2.3.	B.3.1.	B.4.3.	100
#3DEGT	FR67JU8JG	B.1.2.	B.2.3.	B.3.1.	B.4.3.	100
#3DEGT	XCFGBG45	B.1.2.	B.2.3.	B.3.1.	B.4.3.	100
#3DEGT	GDHJUYT5	B.1.2.	B.2.3.	B.3.1.	B.4.3.	100
EV45FR G	RT1H5YEH	B.1.2.	B.2.3.	B.3.2.	B.4.3.	25
EV45FR G	ERT89THY	B.1.2.	B.2.3.	B.3.1.	B.4.1.	0
EV45FR G	YU67TUYT	B.1.2.	B.2.3.	B.3.1.	B.4.3.	100

Table V

Example of average percentage discrimination for entities in a single test

Entity	% Discrimination Settlement
EV45FRG	83.3
#3DEGT	100.0
*3LAV	12.5
943/2M	85.0
48-FS	72.2
1G73KL	100.0
DOKARA	50.0
LIMYRA	87.5

If the number of correct answers meets or exceeds this threshold, it demonstrates statistical evidence of a significant difference at the specified significance level.

Compound identification

Compound identification was carried out according to the possible compounds shown in Table I, which were presented to the evaluators beforehand. An example of the assessor's results concerning sensory compound identification is shown in Table VII. Once the results obtained in terms of compound identification by assessors have been analyzed and compiled, the global percentage of compounds identified was determined (Table VIII), as well as the most frequent value (mode) in terms of compound identification by an entity (Table IX), which is marked in red when the entity was not aligned with the identification of the compound used in the test preparation.

Table VI

Minimum (critical) number of correct answers for a 5% significance level

Different Sample	No. of correct answers	Result[1]	% Correct answers
B.1.2.	188	There is a difference between the samples	87
B.2.3.	196	There is a difference between the samples	90
B.3.1.	198	There is a difference between the samples	91
B.4.3.	203	There is a difference between the samples	94
Minimum number of correct answers for a 5% significance level			85

[1] Green colour: There is a difference between samples

Table VII

Example of assessor's results concerning sensory compound identification

COD_ENTITY	ID_ASSESSOR	COMPOUND
#3DEGT	GT567SHY1	Geosmin
#3DEGT	FR67JU8JG	Geosmin
#3DEGT	XCFGBG45	Geosmin
#3DEGT	GDHJUYT5	Geosmin
EV45FRG	RT1H5YEH	2-Methylisoborneol
EV45FRG	ERT89THY	2-Methylisoborneol
EV45FRG	YU67TUYT	2.4.6-Trichloroanisol

Table VIII

Percentage of assessors identifying each compound in sample B

Sample B	%
Geosmin	46.1
2-Methylisoborneol	22.1
2,4,6-Trichloroanisol	22.1
Not identified	4.1
2-Phenylethanol	1.4
2-Mercaptoethanol	0.9
Guaiacol	0.9
Linalool	0.9
1-Hexanol	0.5
Vanillin	0.5
4-Vinylphenol	0.5

Although the triangle test has traditionally been a widely used difference test, in practice, it is often applied with fewer evaluators able to detect differences if they exist. The triangle test is only applied when the products do not cause excessive sensory fatigue, saturation, and/or adaptation. Other significant limitations include memory effects (delay and memory interference) and psychological factors

(such as error of central tendency), which can affect the performance of the evaluator and the test results (Ojeh, 2020).

Table IX

Modal value panel for compound identification.

Entity	Most frequently (mode) identified compound
EV45FRG	2-Methylisoborneol
#3DEGT	Geosmin
*3LAV	2.4.6-Trichloroanisol
943/2M	2.4.6-Trichloroanisol
48-FS	Geosmin
1G73KL	Geosmin
DOKARA	Geosmin
LIMYRA	2-Methylisoborneol
VENARI	Geosmin
XOCPE	2.4.6-Trichloroanisol

Table X reveals a general improvement in average performance over the years (e.g., 2009–2022). Also, it is possible to observe how the percentage of correct responses (100%, 75%, 50%, 25%, 0%) at different thresholds has changed over the years. For example, "1-Hexanol" saw a significant increase in the percentage of correct responses at the 100% threshold from 2012 to 2016, indicating an improvement in detection or identification accuracy. Increased familiarity with "1-Hexanol" through repeated exposure or a different or more experienced pool of participants in 2016 compared to 2012 could explain the improved detection rates. The "Yearly Change" column reflects each compound's yearly performance. Positive values indicate improvement from one year to the next, while negative values indicate a decline.

Table X

Trends in “correct responses”, yearly performance, and variation across different concentration levels for each compound

Compound	Year	Mean (% correct results)					Yearly Change				
		100[1]	75[2]	50[3]	25[4]	0[5]	100	75	50	25	0
1-Hexanol	2009	26.0	22.0	22.0	11.0	20.0					
	2012	31.0	27.0	13.0	24.0	4.0	5.0	5.0	-9.0	13.0	-16.0
	2016	50.1	16.5	12.3	9.3	4.3	19.1	-10.5	-0.7	-14.7	0.3
2,4,6-Trichloroanisol	2013	30.1	10.0	17.7	21.4	20.9					
	2014	30.9	12.0	13.3	23.5	20.8	0.8	2.1	-4.4	2.1	-0.1
	2015	48.2	9.2	9.2	17.5	16.5	17.4	-2.9	-4.1	-6.0	-4.3
	2016	59.0	10.0	2.0	24.0	4.0	10.8	0.8	-7.2	6.5	-12.5
	2017	59.8	7.6	12.3	11.3	8.8	0.8	-2.4	10.3	-12.7	4.8
	2018	43.0	12.0	13.5	13.5	18.0	-16.8	4.4	1.2	2.2	9.2
	2019	50.0	12.0	11.0	8.0	19.0	7.0	0.0	-2.5	-5.5	1.0
	2020	43.0	11.0	13.0	18.0	15.0	-7.0	-1.0	2.0	10.0	-4.0
	2021	60.0	7.0	7.0	14.0	13.0	17.0	-4.0	-6.0	-4.0	-2.0
	2022	49.0	11.0	8.0	16.0	16.0	-11.0	4.0	1.0	2.0	3.0
2-Phenylethanol	2009	14.0	29.0	10.0	11.0	36.0					
	2011	60.0	7.0	2.0	7.0	24.0	46.0	-22.0	-8.0	-4.0	-12.0
	2013	39.0	18.0	6.0	20.0	17.0	-21.0	11.0	4.0	13.0	-7.0
2-Mercaptoethanol	2010	63.0	4.0	14.0	4.0	14.0					
	2011	28.1	21.8	14.8	21.0	14.3	-34.9	17.8	0.8	17.0	0.3
	2013	67.0	8.0	10.0	6.0	10.0	38.9	-13.8	-4.8	-15.0	-4.3
	2014	12.0	28.0	27.0	19.0	14.0	-55.0	20.0	17.0	13.0	4.0
	2015	15.1	20.4	21.9	22.9	19.5	3.1	-7.6	-5.1	3.9	5.5
	2016	27.0	16.0	15.0	22.0	20.0	11.9	-4.4	-6.9	-0.9	0.5
	2017	19.0	16.0	22.0	15.0	27.0	-8.0	0.0	7.0	-7.0	7.0
	2018	24.0	26.0	23.0	11.0	16.0	5.0	10.0	1.0	-4.0	-11.0
	2021	8.0	16.0	46.0	7.0	24.0	-16.0	-10.0	23.0	-4.0	8.0

[1] An average of 100% indicates that the assessor identified the different sample in all four sequences.
[2] An average of 75% indicates that the assessor identified the different sample in three of the four sequences.
[3] An average of 50% indicates that the assessor identified the different sample in two of the four sequences.
[4] An average of 25% indicates that the assessor identified the different sample in one of the four sequences.
[5] An average of 0% indicates that the assessor did not identify the different sample in any of the four sequences.

Table X (continuation)

Compound	Year	Mean (% correct results)					Yearly Change				
		100[1]	**75**[2]	**50**[3]	**25**[4]	**0**[5]	**100**	**75**	**50**	**25**	**0**
2-Methylisoborneol	2013	27.1	16.7	21.8	18.8	15.4					
	2014	60.0	9.0	8.0	14.0	10.0	32.9	-7.7	-13.8	-4.8	-5.4
	2015	39.0	18.0	29.0	9.0	4.0	-21.0	9.0	21.0	-5.0	-6.0
	2016	42.0	15.0	18.0	7.0	17.0	3.0	-3.0	-11.0	-2.0	13.0
	2017	27.0	13.0	38.0	11.0	11.0	-15.0	-2.0	20.0	4.0	-6.0
	2018	37.0	25.0	8.0	13.0	18.0	10.0	12.0	-30.0	2.0	7.0
	2019	67.9	9.0	12.0	8.0	4.1	30.9	-16.0	4.0	-5.0	-13.9
	2021	20.0	14.0	22.0	13.0	17.0	-47.9	5.0	10.0	5.0	12.9
	2022	10.0	35.0	17.0	18.0	20.0	-10.0	21.0	-5.0	5.0	3.0
2,6 – Lutidine	2010	31.0	6.0	28.0	19.0	17.0					
	2012	64.0	5.0	6.0	4.0	21.0	33.0	-1.0	-22.0	-15.0	4.0
4-Vinylphenol	2012	84.0	6.0	1.0	3.0	5.0					
2-Octanol	2010	85.0	2.0	4.0	1.0	9.0					
	2012	20.0	26.0	17.0	18.0	19.0	-65.0	24.0	13.0	17.0	10.0
4-Ethylphenol	2011	69.3	20.8	2.6	3.0	3.3					
	2014	66.2	15.8	9.3	4.6	4.1	-3.1	-5.0	6.7	1.6	0.8
	2015	76.0	8.0	9.0	5.0	2.0	9.8	-7.8	-0.3	0.4	-2.1
	2016	67.0	18.0	7.0	5.0	3.0	-9.0	10.0	-2.0	0.0	1.0
	2017	63.6	18.0	6.5	5.5	7.4	-3.4	0.0	-0.5	0.5	4.4
	2018	27.0	34.0	22.0	12.0	5.0	-36.6	16.0	15.5	6.5	-2.4
	2019	68.0	9.0	12.0	8.0	4.0	41.0	-25.0	-10.0	-4.0	-1.0
	2021	16.0	19.0	43.0	14.0	8.0	-52.0	10.0	31.0	6.0	4.0
	2022	74.0	10.0	7.0	2.0	8.0	58.0	-9.0	-36.0	-12.0	0.0
Acetaldehyde	2010	39.0	4.0	20.0	6.0	30.0					
Ethyl acetate	2009	8.0	15.0	17.0	20.0	40.0					
Isoamyl acetate	2009	4.5	6.5	11.4	27.5	50.1					
Benzaldehyde	2009	18.4	19.6	21.1	24.3	16.7					
	2010	26.8	21.7	7.8	23.7	19.8	8.4	2.1	-13.3	-0.6	3.1
	2017	48.0	28.0	5.0	5.0	14.0	21.2	6.3	-2.8	-18.7	-5.8
	2020	39.0	22.0	13.0	10.0	16.0	-9.0	-6.0	8.0	5.0	2.0
Citral	2010	10.0	6.0	23.0	16.0	45.0					
	2012	50.0	14.0	22.0	7.0	7.0	40.0	8.0	-1.0	-9.0	-38.0
	2014	24.0	5.0	19.0	25.0	26.0	-26.0	-9.0	-3.0	18.0	19.0
Coumarin	2016	33.0	22.0	31.0	3.0	12.0					
	2017	46.0	20.0	18.0	9.0	7.0	13.0	-2.0	-13.0	6.0	-5.0
	2021	19.0	22.0	27.0	22.0	12.0	-27.0	2.0	9.0	13.0	5.0
Diacetyl	2010	12.0	8.0	20.0	43.0	16.0					
	2012	32.0	28.0	18.0	11.0	11.0	20.0	20.0	-2.0	-32.0	-5.0
	2015	40.0	26.0	18.0	8.0	8.0	8.0	-2.0	0.0	-3.0	-3.0
	2016	31.0	7.0	35.0	7.0	19.0	-9.0	-19.0	17.0	-1.0	11.0
Eucalyptol	2010	23.0	6.0	21.0	12.0	38.0					
	2011	51.0	4.0	9.0	7.0	29.0	28.0	-2.0	-12.0	-5.0	-9.0
	2012	61.0	6.0	16.0	3.0	14.0	10.0	2.0	7.0	-4.0	-15.0
	2014	66.0	4.0	5.0	17.0	8.0	5.0	-2.0	-11.0	14.0	-6.0
Eugenol	2009	17.4	26.1	11.7	13.5	30.8					
	2014	63.0	12.0	5.0	6.0	12.0	45.6	-14.1	-6.7	-7.5	-18.8

[1] An average of 100% indicates that the assessor identified the different sample in all four sequences.
[2] An average of 75% indicates that the assessor identified the different sample in three of the four sequences.
[3] An average of 50% indicates that the assessor identified the different sample in two of the four sequences.
[4] An average of 25% indicates that the assessor identified the different sample in one of the four sequences.
[5] An average of 0% indicates that the assessor did not identify the different sample in any of the four sequences.

Table X (continuation)

Compound	Year	Mean (% correct results)					Yearly Change				
		100[1]	**75**[2]	**50**[3]	**25**[4]	**0**[5]	**100**	**75**	**50**	**25**	**0**
Fenchone	2012	37.0	11.0	9.0	21.0	22.0					
Guaiacol	2009	10.0	18.0	26.0	17.0	29.0					
	2011	53.0	11.0	21.0	8.0	7.0	43.0	-7.0	-5.0	-9.0	-22.0
	2013	35.3	15.9	31.0	8.4	9.7	-17.7	4.9	10	0.4	2.7
	2014	33.6	23.1	30.0	6.2	7.1	-1.7	7.2	-1.1	-2.2	-2.6
	2015	48.0	24.0	16.0	8.0	3.0	14.4	0.9	-14.0	1.8	-4.1
	2016	13.0	15.0	37.0	21.0	13.0	-35.0	-9.0	21.0	13.0	10.0
	2018	45.0	20.0	16.0	14.0	6.0	32.0	5.0	-21.0	-7.0	-7.0
	2019	42.0	18.0	15.0	16.0	8.0	-3.0	-2.0	-1.0	2.0	2.0
	2021	31.6	6.4	9.4	9.2	42.9	-10.4	-11.6	-5.6	-6.8	34.9
	2022	83.0	1.0	10.0	3.0	4.0	51.4	-5.4	0.6	-6.2	-38.9
Limonene	2010	37.0	4.0	3.0	38.0	18.0					
	2011	51.0	7.0	15.0	1.0	26.0	14.0	3.0	12.0	-37.0	8.0
Linalool	2010	28.0	14.0	3.0	38.0	16.0					
	2012	45.4	9.8	15.6	10.9	18.8	17.4	-4.2	12.6	-27.1	2.8
	2015	64.0	7.0	12.0	3.0	14.0	18.6	-2.8	-3.6	-7.9	-4.8
	2017	53.0	11.0	12.0	13.0	12.0	-11.0	4.0	0.0	10.0	-2.0
	2019	25.0	9.0	11.0	7.0	48.0	-28.0	-2.0	-1.0	-6.0	36.0
	2022	51.0	18.0	7.0	9.0	15.0	26.0	9.0	-4.0	2.0	-33.0
Pyridine	2011	18.0	21.0	9.0	22.0	29.0					
Vanillin	2009	6.0	9.0	15.0	24.0	47.0					

[1] An average of 100% indicates that the assessor identified the different sample in all four sequences.
[2] An average of 75% indicates that the assessor identified the different sample in three of the four sequences.
[3] An average of 50% indicates that the assessor identified the different sample in two of the four sequences.
[4] An average of 25% indicates that the assessor identified the different sample in one of the four sequences.
[5] An average of 0% indicates that the assessor did not identify the different sample in any of the four sequences.

These changes can be used to track the consistency of performance for each compound over the years. By comparing the changes across different thresholds, it is possible to identify which levels of certainty (e.g., 100% versus 50%) experienced the most notable shifts. These results reveal how difficult it was for respondents to identify each compound correctly at varying confidence levels. This analysis is intended to enhance the understanding of the trends in identification accuracy and the relative difficulty associated with each compound over time. Compounds tested for the first time, such as in 2009 and 2010, often show lower performance than subsequent years. This improvement in later years may be assigned to the benefits of consistent testing ranges.

The analysis of Table XI will help understanding how the concentration and performance scores have changed over the years.

Compounds such as 4-vinylphenol (90.2%, 2012) and linalool (75.8%, 2015) showed high average ratings, likely to indicate familiarity or straightforward detection. Compounds like diacetyl (39.3%, 2010) and isoamyl acetate (18.8%, 2009) had lower ratings, potentially due to challenges in perception or detection limits. For certain compounds like guaiacol and 2-mercaptoethanol, performance ratings improved significantly over multiple testing years, reflecting enhanced detection skills. High variance in performance for compounds like coumarin and guaiacol suggests challenges at extreme concentration levels. Years with higher average ratings in Table X (e.g., 2016, 2017) align with improved compound-specific ratings for commonly tested or critical compounds such as 2,4,6-trichloroanisol, 4-ethylphenol, and linalool. Consistent testing and training likely led to better overall performance. Introducing new or less familiar compounds (e.g., fenchone and coumarin) corresponded to performance variability. Training and familiarity with these compounds in subsequent years helped stabilize the ratings.

Table XI

Compounds and concentrations used in each of the four triangle test sequences in different editions of different years

Compound	Units	Year	Edition	Conc.1	Conc.2	Conc.3	Conc.4	% Average final performance rating [0 - 100]
1-Hexanol	mg/L	2009	3	1.0	2.0	3.0	4.5	55.5
	mg/L	2012	1	1.0	1.6	2.6	4.1	63.9
	mg/L	2016	1	1.4	2.2	3.8	5.4	86.6
	mg/L	2016	2	1.1	1.8	3.4	5.0	78.8
	mg/L	2016	3	1.1	1.8	3.4	5.0	76.0
	mg/L	2016	4	1.1	1.8	3.4	5.0	73.7
2,4,6-Trichloroanisol	ng/L	2013	1	1.2	1.8	2.4	3.0	49.3
	ng/L	2013	2	1.2	1.8	2.7	4.0	59.2
	ng/L	2013	3	1.2	1.8	2.7	4.1	58.9
	ng/L	2013	4	1.2	1.8	2.7	4.0	41.3
	ng/L	2014	1	1.2	1.8	2.7	4.0	54.2
	ng/L	2014	4	1.2	1.8	2.7	4.0	50.4
	ng/L	2015	2	1.2	1.8	2.7	4.1	65.2
	ng/L	2015	4	1.2	1.8	2.7	4.0	63.0
	ng/L	2016	1	1.2	1.8	2.7	4.1	73.7
	ng/L	2017	1	1.2	1.8	2.7	4.1	74.3
	ng/L	2017	3	1.2	1.8	2.7	4.1	74.6
	ng/L	2018	1	1.2	1.8	2.7	4.0	62.3
	ng/L	2018	3	1.2	1.8	2.7	4.1	62.2
	ng/L	2019	1	1.2	1.8	2.7	4.0	66.7
	ng/L	2020	3	1.2	1.8	2.7	4.0	62.5
	ng/L	2021	3	1.4	2.1	3.2	4.5	71.9
	ng/L	2022	2	1.4	2.1	3.2	4.5	65.3
2,6 – Lutidine	mg/L	2010	4	0.8	1.0	1.2	1.4	53.8
	mg/L	2012	4	1.0	1.2	1.6	2.1	71.9
2-Phenylethanol	mg/L	2009	1	0.3	0.5	0.8	1.2	44.0
	mg/L	2011	4	0.3	0.5	0.8	1.2	67.9
	mg/L	2013	4	0.3	0.5	0.8	1.2	60.8
2-Mercaptoethanol	mg/L	2010	3	1.3	1.9	2.9	3.9	74.4
	mg/L	2011	1	1.0	1.6	2.2	2.8	64.2
	mg/L	2011	2	1.0	1.6	2.2	2.8	56.6
	mg/L	2011	3	1.0	1.6	2.2	2.8	63.9
	mg/L	2011	4	1.0	1.6	2.2	2.8	42.4
	mg/L	2013	2	1.0	1.6	2.2	2.8	79.6
	mg/L	2014	3	0.2	0.6	1.0	1.4	51.7
	mg/L	2015	1	0.2	0.6	1.0	1.4	53.0
	mg/L	2015	2	0.2	0.6	1.2	1.8	51.4
	mg/L	2015	3	0.2	0.6	1.0	1.4	50.5
	mg/L	2015	4	0.2	0.6	1.0	1.4	36.2
	mg/L	2016	2	0.3	0.7	1.1	1.5	51.6
	mg/L	2017	3	0.3	0.7	1.1	1.5	46.3
	mg/L	2018	2	0.4	0.8	1.3	1.7	57.6
	mg/L	2021	3	0.6	1.1	1.9	3.4	44.2
4-Vinylphenol	mg/L	2012	3	0.30	0.36	0.43	0.52	90.2

Table XI (continuation)

Compound	Units	Year	Edition	Conc.1	Conc.2	Conc.3	Conc.4	% Average final performance rating [0 - 100]
	ng/L	2013	1	3.2	8.2	13.2	18.2	54.1
	ng/L	2013	2	2.8	6.2	13.5	29.8	54.9
	ng/L	2013	4	2.8	6.2	13.5	29.8	55.9
	ng/L	2014	2	2.8	6.2	13.6	29.8	73.6
	ng/L	2015	1	2.1	4.6	10.2	22.4	69.6
2-Methylisoborneol	ng/L	2016	2	2.1	4.6	10.2	22.4	64.7
	ng/L	2017	4	2.1	4.6	10.2	22.4	58.6
	ng/L	2018	2	2.6	5.8	12.6	28.0	62.7
	ng/L	2019	2	2.6	5.8	12.6	28.0	61.4
	ng/L	2021	1	2.6	5.8	12.8	27.7	48.9
	ng/L	2022	1	2.6	5.8	12.6	28.0	49.5
2-Octanol	mg/L	2010	4	0.50	0.70	0.90	1.40	88.0
	mg/L	2012	4	0.50	0.13	0.31	0.78	52.6
	mg/L	2011	1	0.10	0.40	0.70	1.00	79.6
	mg/L	2011	3	0.05	0.35	0.65	0.95	91.5
	mg/L	2011	4	0.05	0.15	0.25	0.35	91.1
	mg/L	2014	1	0.04	0.08	0.16	0.32	87.1
	mg/L	2014	2	0.03	0.06	0.11	0.22	88.8
	mg/L	2014	3	0.02	0.05	0.09	0.18	79.9
	mg/L	2014	4	0.02	0.05	0.09	0.18	80.6
4-Ethylphenol	mg/L	2015	3	0.02	0.05	0.09	0.18	87.6
	mg/L	2016	1	0.02	0.05	0.09	0.18	85.4
	mg/L	2017	1	0.02	0.04	0.07	0.09	75.7
	mg/L	2017	4	0.02	0.05	0.09	0.18	86.4
	mg/L	2018	1	0.02	0.04	0.08	0.16	66.6
	mg/L	2019	2	0.02	0.04	0.08	0.16	81.4
	mg/L	2021	2	0.04	0.08	0.16	0.32	55.3
	mg/L	2022	3	0.04	0.08	0.16	0.32	85.3
Acetaldehyde	mg/L	2010	2	0.14	0.19	0.24	0.29	54.1
Ethyl acetate	mg/L	2009	1	3	5	8	15	32.4
Isoamyl acetate	mg/L	2009	1	0.008	0.02	0.04	0.10	18.8
	mg/L	2009	2	0.02	0.04	0.10	0.30	25.0
	mg/L	2009	2	0.4	0.6	0.8	1.0	48.8
	mg/L	2009	4	0.4	0.6	0.8	1.0	50.2
	mg/L	2010	1	0.4	0.6	0.8	1.0	61.2
Benzaldehyde	mg/L	2010	2	0.4	0.6	0.8	1.0	27.0
	mg/L	2010	3	0.4	0.6	0.8	1.0	56.0
	mg/L	2010	4	0.4	0.6	0.8	1.0	68.8
	mg/L	2017	2	0.4	0.6	0.8	1.0	72.6
	mg/L	2020	3	0.4	0.6	0.8	1.0	64.3
Citral	mg/L	2010	1	0.09	0.15	0.23	0.34	29.7
	mg/L	2012	2	0.12	0.19	0.31	0.49	73.5
	mg/L	2014	4	0.12	0.19	0.31	0.49	44.2
Coumarin	µg/L	2016	4	5	15	45	135	65.3
	µg/L	2017	3	5	15	45	135	72.4
	µg/L	2021	1	5	15	45	135	53.4
Eucalyptol	mg/L	2010	1	0.008	0.010	0.012	0.014	41.2
	mg/L	2011	2	0.009	0.013	0.017	0.020	60.5
	mg/L	2012	1	0.010	0.014	0.020	0.027	74.1
	mg/L	2014	1	0.010	0.014	0.020	0.027	75.8
Eugenol	mg/L	2009	2	0.001	0.004	0.008	0.020	47.6
	mg/L	2009	3	0.001	0.004	0.008	0.020	44.8
	mg/L	2014	2	0.001	0.005	0.016	0.056	77.0
Fenchone	mg/L	2012	2	0.52	0.62	0.74	0.89	55.4

Table XI (continuation)

Compound	Units	Year	Edition	Conc.1	Conc.2	Conc.3	Conc.4	% Average final performance rating [0 - 100]
Guaiacol	µg/L	2009	3	5	10	30	50	40.5
	µg/L	2011	3	5	10	30	50	73.8
	µg/L	2013	2	0.3	0.8	2.3	6.8	63.4
	µg/L	2013	3	0.4	1.1	3.1	9.5	72.0
	µg/L	2013	4	0.4	1.1	3.1	9.5	59.5
	µg/L	2014	3	0.4	1.1	3.2	9.5	67.6
	µg/L	2015	3	0.4	1.1	3.2	9.5	76.8
	µg/L	2016	4	0.4	1.1	3.1	9.5	48.5
	µg/L	2018	3	0.4	1.3	3.8	11.3	71.0
	µg/L	2019	3	0.5	1.5	4.5	13.7	67.4
	µg/L	2021	2	1.5	4.5	13.5	40.5	6.7
	µg/L	2021	3	1.5	4.5	13.5	40.5	75.7
	µg/L	2022	3	4.5	11.3	28.1	70.3	88.8
Limonene	mg/L	2010	2	0.13	0.18	0.23	0.28	50.8
	mg/L	2011	1	0.15	0.22	0.27	0.32	64.1
Linalool	mg/L	2010	3	0.01	0.02	0.05	0.08	50.0
	mg/L	2012	1	0.01	0.02	0.05	0.10	61.9
	mg/L	2012	2	0.01	0.02	0.05	0.10	64.1
	mg/L	2012	3	0.01	0.02	0.05	0.10	67.1
	mg/L	2012	4	0.01	0.02	0.05	0.10	59.3
	mg/L	2015	2	0.01	0.02	0.05	0.10	75.8
	mg/L	2017	4	0.01	0.02	0.05	0.10	70.0
	mg/L	2019	3	0.01	0.02	0.05	0.10	38.7
	mg/L	2022	1	0.01	0.02	0.05	0.10	69.9
Vanillin	mg/L	2009	4	0.06	0.09	0.12	0.20	25.8
Diacetyl	mg/L	2010	1	0.005	0.006	0.010	0.030	39.3
	mg/L	2012	3	0.002	0.004	0.008	0.016	65.2
	mg/L	2015	1	0.002	0.004	0.008	0.016	70.8
	mg/L	2016	3	0.002	0.004	0.008	0.016	55.8
Pyridine	mg/L	2011	2	2	3	4	5	43.9

The observation of Tables X and XI could provide some additional insights. Additional training for compounds with historically low ratings, such as diacetyl, isoamyl acetate, and pyridine, could enhance overall performance. Stabilizing the test concentration ranges could reduce variability and improve performance ratings.

Triangle tests with lower concentrations (Conc.1 and 2) seemed to have relatively consistent trends for most compounds, with some variation depending on the compound and year. In contrast, higher concentrations (Conc.3 and 4) showed more variability, with certain compounds exhibiting distinct trends, either increasing or decreasing over time.

Slight changes in the concentration of the compounds were observed over the years and across different editions. These adjustments were made based on studies of the population's behavior and/or the need to address issues such as the test being too easy or too difficult. The goal was to ensure balanced and accurate results. This change in concentration, as well as variations in assessors (and entities) between years/editions, somewhat hinders the ability to compare results across different editions and years. The population is dynamic and performs on an annual basis, and the variation between the various concentration levels of each compound must be considered. However, given that the final performance results for each compound derived from a substantial population of assessors - typically with a median of 164, ranging from a minimum of 84 to a maximum of 235 - there is reasonable confidence in the values reported. In general, an increase in the concentration of a given compound leads to an increase in the percentage of correct answers, regardless of the possible variations that may occur.

The frequency of assessment of each compound varied, and was much higher for some compounds due to their relevance to the activities of most of the member organizations. The compounds used and their assessment, together with data concerning compound performance trend, are shown in Table XII. The top performers' compounds are guaiacol (+119.3%) and eucalyptol (+83.8%), which showed the highest percentage increases in performance. Eugenol (+66.7%) also demonstrated a solid upward trend.

Table XII

Ranked compounds by change (%) with evaluation frequency of each compound

Compound	Evaluation Frequency	Initial Year	Final Year	Initial Value	Final Value	Trend	Value Change	Change (%)
Guaiacol	13	2009	2022	40.5	88.8	Increase	48.3	119.3
Eucalyptol	4	2010	2014	41.2	75.8	Increase	34.6	83.8
Eugenol	3	2009	2014	46.2	77.0	Increase	30.8	66.7
Citral	3	2010	2014	29.7	44.2	Increase	14.5	48.8
1-Hexanol	6	2009	2016	55.5	78.8	Increase	23.3	42.1
Diacetyl	4	2010	2016	39.3	55.8	Increase	16.5	41.8
Linalool	9	2010	2022	50.0	69.9	Increase	19.9	39.9
2-Phenylethanol	3	2009	2013	44.0	60.8	Increase	16.8	38.2
2,6 - Lutidine	2	2010	2012	53.8	71.9	Increase	18.1	33.7
Benzaldehyde	8	2009	2020	49.5	64.3	Increase	14.8	30.0
Limonene	2	2010	2011	50.8	64.1	Increase	13.3	26.1
2,4,6-Trichloroanisol	17	2013	2022	52.2	65.3	Increase	13.2	25.2
Vanillin	1	2009	2009	25.8	25.8	No Change	0.0	0.0
Fenchone	1	2012	2012	55.4	55.4	No Change	0.0	0.0
Isoamyl acetate	2	2009	2009	21.9	21.9	No Change	0.0	0.0
Ethyl acetate	1	2009	2009	32.4	32.4	No Change	0.0	0.0
Acetaldehyde	1	2010	2010	54.1	54.1	No Change	0.0	0.0
4-Vinylphenol	1	2012	2012	90.2	90.2	No Change	0.0	0.0
Pyridine	1	2011	2011	43.9	43.9	No Change	0.0	0.0
4-Ethylphenol	15	2011	2022	87.4	85.3	Decrease	-2.1	-2.4
2-Methylisoborneol	11	2013	2022	55.0	49.5	Decrease	-5.5	-10.0
Coumarin	3	2016	2021	65.3	53.4	Decrease	-11.9	-18.2
2-Octanol	2	2010	2012	88.0	52.6	Decrease	-35.4	-40.2
2-Mercaptoethanol	15	2010	2021	74.4	44.2	Decrease	-30.2	-40.6

Some compounds showed some relevant performance declines; 2-Mercaptoethanol (-40.6%) and 2-octanol (-40.2%) exhibited the most significant decreases. The performance of 2-mercaptoethanol most likely declined due to a decrease in concentration values, particularly between the years 2014 and 2018. The decrease in performance of 2-octanol was expected due to the decrease in concentration values. Coumarin (-18.2%) experienced a significant decline. Some compounds, such as ethyl acetate (0%), showed no change in trend due to minimal frequency evaluation.

The study compared the impact of adding a compound to only one sample (1 positive) versus adding it to two samples (2 positive). The results were analyzed to determine whether there were any significant differences in participants' perceptions between these two conditions. Considering these assumptions, the necessary information was gathered to determine whether differences in performance existed, as shown in Table XIII, Figure 1 and Figure 2.

Table XIII

Effect one (1) positive vs two (2) positive in a triangle test

Sequence	Positive samples	Settlement[1]	Counting	Overall count	%
Seq. 1	1	1	6399	11093	57.7
	2	1	4694		42.3
Seq. 1	1	0	3937	8087	48.7
	2	0	4150		51.3
Seq. 2	1	1	6476	12975	49.9
	2	1	6499		50.1
Seq. 2	1	0	2890	6217	46.5
	2	0	3327		53.5
Seq. 3	1	1	7353	14909	49.3
	2	1	7556		50.7
Seq. 3	1	0	1479	4283	34.5
	2	0	2804		65.5
Seq. 4	1	1	8505	16053	53.0
	2	1	7548		47.0
Seq. 4	1	0	1332	3139	42.4
	2	0	1807		57.6

[1] 1 = get it right; 0 = can not get it right.

The data indicates that the percentage of correct answers in cases with only one positive sample is significantly higher for sequences 1 (lowest concentration) and 4 (highest concentration). On the other hand, the percentage of incorrect answers is always higher in cases where there are two positive samples, being mostly higher in concentration sequence three and slightly higher in concentration sequence four when compared to concentration sequences 1 and 2. This aligns with the findings of several researchers (Hopkins, 1954; Frijters, 1977; O'Mahony, 1995) who demonstrated that a triad containing a strong stimulus as the odd sample is easier to discriminate compared to one with a weak stimulus as the odd sample. O'Mahony (1995) also stated that the triad's first stimulus might have been influenced by other factors, such as a mouth rinse between the triads or the previous triad's final

stimulus. Even when considering these factors, the prediction for the triads remains unchanged.

Filipello (1956) ascribed this effect to decreased sensitivity due to sensory adaptation during a tri-stimulus sequence.

There are some publications on triangle tests with replicates (Frijters *et al.*, 1982; Kunert and Meyners, 1999). Kunert and Meyners (1999) highlighted the significant negative impact of assessor heterogeneity on the statistical power of the naïve binomial test when aiming to demonstrate product differences.

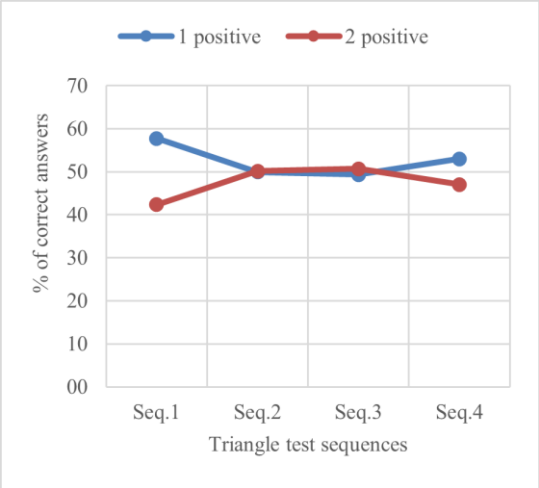

Figure 1. Effect 1 positive vs 2 positive. % of evaluation correct answers.

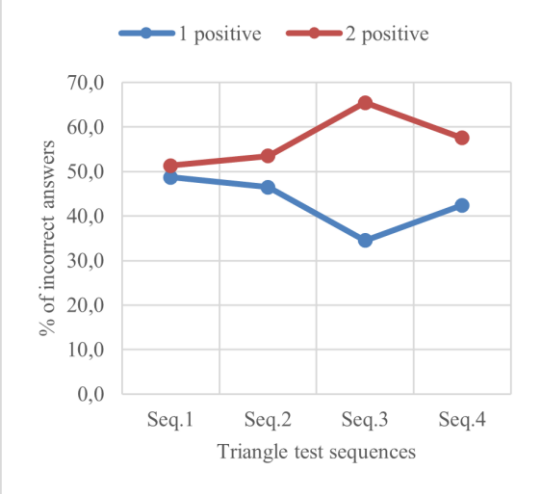

Figure 2. Effect 1 positive vs 2 positive. % of evaluation incorrect answers.

They illustrate this with an example, where in case 1, a single assessor provides 100 answers. There is a 50% probability the assessor is "good," resulting in 100 correct answers, and a 50% probability the assessor is "poor," leading to a binomial distribution with parameters m = 100 and p = 1/3. The probability of obtaining a significant result at the 5% level is 1 if the assessor is "good" and 0.05 if "poor," yielding an overall probability of 0.525 for rejecting the null hypothesis. In case 2, 100 assessors each provide one answer. Each assessor has a 50% chance of being "good" (correct with probability 1) or "poor" (correct with probability 1/3). Overall, each answer has a 2/3 probability of being correct, resulting in a binomial distribution with parameters m = 100 and p = 2/3. The probability of observing more than 42 correct answers (the critical value for a triangle test with 100 assessments) exceeds 99%, meaning the null hypothesis is almost certain to be rejected. Case 2, with 100 independent assessors, is far more reliable for rejecting the null hypothesis (probability > 99%) compared to case 1, which relies on a single assessor and has only a 52.5% probability of achieving significance. Distributing assessments across multiple individuals significantly improves the robustness and reliability of the results. The lower power in the latter case is due to the risk that the single assessor might be insensitive and, therefore, merely guessing. If all assessors have the same true success rate, there is no heterogeneity (sometimes referred to as assessor homogeneity instead); that is, assessor heterogeneity in this context means that assessors vary in their sensitivity and, hence, in their true success probabilities for the (discrimination) task repeatedly presented to them (Meyners and Carr, 2024).

However, SENSORIAL-ALABE cannot be considered to have true replicates, as each of the four sequences had a different concentration of the same compound. The results generally show an improvement in the performance of the assessor's answers when laboratories are faced with carrying out the test for a second time, which is in line with similar studies (McKay *et al.*, 2018). It is also worth considering the hypothesis of different perception thresholds between assessors or even specific anosmia by part of the population for some of the compounds presented, which was evidenced in various publications in the field, such as those referred to in the study by Gaby *et al.* (2020).

Exceptional assessors offer significant benefits across various industries, as Pinto *et al.* (2024) have already described. Based on the discrimination performance, the overall performance indicator was calculated for the organizations and the assessors, showing performances ranging from 0.0% to 1759.9% for organizations, and 0% to 178.3% for assessors.

CONCLUSIONS

Despite the quantities added *versus* the effective concentrations of the added compounds, the results obtained in the proficiency test show an improvement in discrimination capacity and

consequent improvement in the % correct with repeated exposure to the added compound. Although the time between tests is long (around four months) and the concentrations of the analytes are not always the same between tests, assessors generally discriminate better between artificially added compounds if the molecule responsible for the change has already been assessed in previous tests. The average final performance rating for 4-ethylphenol increased from 55.3 in the 2021 (2nd edition) to 85.3 in the 2022 (3rd edition), indicating that assessors' performance improves when laboratories conduct the test a second time. These results strongly suggest that the PT is essential for the continuous improvement of laboratories, improves the sensitivity of assessors because of greater training, and contributes to their qualification by providing relevant information on their performance.

When comparing the effect of adding a compound to just one of the samples (1 positive) *versus* adding it to two samples (2 positives), the number and position of the compounds in a triangle test, combined with the concentration and consequent olfactory intensity, influences the subsequent % hit rate, suggesting that it may be easier for assessors to identify due to its distinctiveness, but if it's added to two samples, it can create more complexity, making it harder to discern the correct sample.

The results indicate that the practices conducted effectively helped assessors enhance their ability to discriminate between the aromas of different product samples. Consequently, sensory analysis, as a scientific discipline, played a crucial role in improving the assessors' skills through a structured, educational approach.

ACKNOWLEDGEMENTS

Acknowledgments are due to ALABE for making this study possible. Without the support, this work would not have been viable.

This study was funded by the CQ-VR [grant number UIDB/00616/2020 and UIDP/00616/2020, https://doi.org/10.54499/UIDB/00616/2020] and by the CEMAT/IST-ID [grant number UIDB/04621/2020 and UIDP/04621/2020, https://doi.org/10.54499/UIDB/04621/2020]

CONFLICTS OF INTEREST: The authors declare no conflict of interest.

REFERENCES

ALABE, 2024. Instruções/informações SENSORIAL-ALABE. Associação dos Laboratórios de Enologia,. Available at: https://www.alabe.pt/docs/sensorial/doc_instru%C3%A7%C3%B5es_sa2021_v10.pdf (accessed on 27.12.2024).

ASTM E1885-04, 2011. Standard Test Method for Sensory Analysis - Triangle Test.

AWRI (Australian Wine Research Institute), 2024. Wine flavours, faults and taints. Available at: https://www.awri.com.au/industry_support/winemaking_resources/sensory_assessment/recognition-of-wine-faults-and-taints/wine_faults/ (accessed on 30.08.2024).

Beeren C., 2018. Application of Descriptive Sensory Analysis to Food and Drink Products. *In:Descriptive Analysis in Sensory Evaluation*. 609-646. Kemp S.E., Hort J., Hollowood T. (eds.), Wiley-Blackwell, West Sussex.

Burdock G.A., 2016. Fenaroli's handbook of flavor ingredients. 2159 p. CRC Press, Boca Raton.

Civille G.V., Carr B.T., Osdoba K.E., 2024. Sensory Evaluation Techniques (6 ed.). 999 p. CRC Press. Boca Raton.

Curtis R.F., Land D.G., Griffiths N.M., Gee M., Robinson D., Peel J.L., Dennis C., Gee J.M., 1972. 2,3,4,6-Tetrachloroanisole Association with Musty Taint in Chickens and Microbiological Formation. *Nature*, **235**, 223-224

Diep T., Yoo M., Pook C., Sadooghy-Saraby S., Gite A., Rush E., 2021. Volatile Components and Preliminary Antibacterial Activity of Tamarillo (Solanum betaceum Cav.). *Foods*, **10**, 2212.

Duizer L.M., Walker S.B., 2016. The Application of Sensory Science to the Evaluation of Grain-Based Foods. *In* : *Encyclopedia of Food Grains* (Second Edition). 144-153. Wrigley C., Corke H., Seetharaman K., Faubion J. (Eds.), Academic Press, New York.

EA, 2022. Accreditation for sensory testing laboratories. In: European co-operation for Accreditation, Laboratory Committee. (EA-4/09 G: 2022). European Accreditation, Utrecht.

Filipello F., 1956. A Critical comparison of the two - sample and triangular binomial designs. *J. Food Sci*., **21**, 235-241.

Ford R.A., 2017. Deciding Which Test to Use in Discrimination Testing. *In: Discrimination Testing in Sensory Science: A Practical Handbook*, 67-83. Rogers L. (Ed.) Woodhead Publishing. Sawston.

Frijters J.E.R., 1977. The effect of duration of intervals between olfactory stimuli in the triangular method. *Chem. Senses*, **2**, 301-311.

Frijters J.E.R., Blauw Y.H., Vermaat S.H., 1982. Incidental training in the Triangular Method. *Chem. Senses*, **7**, 63-69.

Gaby J. M., Bakke A.J., Baker A.N., Hopfer H., Hayes J.E., 2020. Individual Differences in Thresholds and Consumer Preferences for Rotundone Added to Red Wine. *Nutrients*, **12**, 2522.

Grainger K., 2021. Wine Faults and Flaws: A Practical Guide. 528 p. Wiley-Blackwell, West Sussex.

Greim H., 2024. 2-Mercaptoethanol. *In : Encyclopedia of Toxicology (4th edition).* 111-115. Wexler P. (Ed.), Academic Press. Boca Raton.

Griffiths N.M., 1974. Sensory properties of the chloro-anisoles. *Chem. Senses*, **1**, 187-195.

Hopkins J.W., 1954. Some Observations on Sensitivity and Repeatability of Triad Taste Difference Tests. *Biometrics*, **10**, 521-530.

Hyldig G., 2010. Proficiency testing of sensory panels. *In : Sensory Analysis for Food and Beverage Quality Control*. 37-48. Kilcast D. (Ed.), Woodhead Publishing, Cambridge.

IPAC., 2019. DRC005 - Procedimento para Acreditação de Laboratórios. Available at: http://www.ipac.pt/docs/publicdocs/regulamentos/DRC005_ProcAcrLabs_v20191106.pdf (accessed on 23.08.2024).

ISO 3591, 1977. Sensory analysis- Apparatus-Wine-tasting glass. International Organization of Standardization. Genève. Switzerland.

ISO/IEC 17025, 2017. General requirements for the competence of testing and calibration laboratories. International Organization of Standardization. Genève. Switzerland.

ISO 4120, 2021 - Sensory analysis — Methodology — Triangle test. International Organization of Standardization. Genève.

ISO/IEC 17043, 2023. Conformity assessment-General requirements for the competence of proficiency testing providers. International Organization of Standardization. Genève. Switzerland.

Kemp, D. S. E., Hollowood, D. T., Hort, D. J., 2009. Sensory Test Methods. *In : Sensory Evaluation - A practical handbook.* 66-137. Wiley-Blackwell, West Sussex.

Kemp S.E., Hollowood T., Hort J., Rogers L..2024. Introduction. *In : Discrimination Testing in Sensory Evaluation*. 1-24. Wiley. West Sussex.

Kilcast D., 2010. Sensory Analysis for Food and Beverage Quality Control: A Practical Guide. 400 p. Woodhead Publishing, Cambridge.

Kunert J., Meyners M., 1999. On the triangle test with replications. *Food Qual. Prefer.*, **10**, 477-482.

Lawless H.T., Heymann H., 2010. Sensory Evaluation of Food: Principles and Practices (2 ed.). 596 p. Springer Verlag. New York.

Mitterer-Daltoé M.L, De Oliveira Treptow R., Martins E., Martins V.M.V., Queiroz M. I., 2012. Selecting and Training a Panel to Evaluate the Metallic Sensation of Meat. *Food Sci Technol Res*., 18(2), 279-286.

Malleret L., Bruchet A., Hennion M.C., 2001. Picogram Determination of "Earthy-Musty" Odorous Compounds in Water Using Modified Closed Loop Stripping Analysis and Large Volume Injection GC/MS. *Anal. Chem.*, **73**, 1485-1490.

McEwan J.A., Heiniö R.L., Hunter E.A., Lea P., 2003. Proficiency testing for sensory ranking panels: measuring panel performance. *Food Qual. Prefer.*, **14**, 247-256

McKay M., Bauer F., Panzeri V., Buica A., 2018. Testing the Sensitivity of Potential Panelists for Wine Taint Compounds Using a Simplified Sensory Strategy. *Foods*, **7**, 176.

Meilgaard M.C., Civille G.V., Carr B.T., 2015. Sensory Evaluation Techniques (5 ed.). 589 p. CRC Press, Boca Raton.

Meyners M., Carr B.T., 2024. Replicated Discrimination Testing. *In: Discrimination Testing in Sensory Evaluation.* 151-195. Rogers L., Hort J., Kemp S.E., Hollowood T. *(eds).* Wiley-Blackwell, West Sussex.

O'Mahony M., 1986. Sensory Evaluation of Food: Statistical Methods and Procedures. 510 p. Marcel Dekker Inc., New York

O'Mahony M., 1995. Who told you the triangle test was simple? *Food Qual. Prefer.*, **6**, 227-238.

Ojeh S., 2020. Discrimination Test Methods. *In: Sensory Testing Methods (3 ed)*. 25-49. West Conshohocken, USA.

Palermo, J.-R., 2015. Análise sensorial: fundamentos e métodos (1 ed). 140 p. Atheneu, São Paulo.

Persson P.-E., 1980. Sensory properties and analysis of two muddy odour compounds, geosmin and 2-methylisoborneol, in water and fish. *Water Res.*, **14**, 1113-1118.

Pinto M., Barros P., Vilela A., Correia E., 2024. Controlling sensory analysis results: tools and importance in assessor selection. Case study: PROVA-ALABE proficiency test. *Ciência Téc. Vitiv*, **39**, 30-50.

Roessler E. B., Pangborn R., Sidel J. L., Stone H., 1978. Expanded statistical tables for estimating significance in paired-preference, paired difference, duo-trio and triangle tests. *J. Food Sci*., **43**, 940-943.

Rogers L., 2017. Appendix 2 - Statistical Tables. *In: Discrimination Testing in Sensory Science*. 363-487. Rogers L. (Ed.), Woodhead Publishing, Cambridge.

Schultz A., 2021. The Basis of Sensory Data for QC: Sensory Specifications. *In: The Role of Sensory Analysis in Quality Control (2 ed).* 47-54. West Conshohocken.

Sinkinson C., 2017. Triangle Test. *In : Discrimination Testing in Sensory Science*. 153-170. Rogers L. (ed.), Woodhead Publishing. Sawston.

Stone H., Bleibaum R.N., Thomas H.A., 2012. Sensory Evaluation Practices (4 ed.). 438 p. Academic Press, Inc., San Diego.

Yeretzian C., 2017. Coffee. *In : Springer Handbook of Odor*. 21-22. Buettner A. (Ed.), Springer International Publishing, New York.