The Australian Wine Research Institute

A procedure for sensory evaluation of wine attributes.

Dr Patricia Williamson, Research Scientist
Dr Leigh Francis, Research Manager

Purpose and scope

A sensory procedure is presented for application by the wine industry for evaluation of wines for the presence of specific attributes, notably taints, off-flavours or other undesirable attributes, and particularly from those arising from the effect of smoke from bushfires on vineyards during the growing season.

The purpose of this document is to provide a guide for grape and wine producers to evaluate wine sensory attributes in an objective matter, following standardised procedures to achieve reliable, repeatable and accurate results. The sensory methodology is provided with validation data for the example of evaluation of vine smoke exposure and resulting wine taint.

The use of good practices in a panel-based blind sensory evaluation ensures decisions will be made based on established and accepted scientific principles, and that the outcomes will be clear, actionable, and reliable.

This procedure has been developed based on knowledge of the fundamentals of sensory science, following a review of small panel and quality control sensory methods, and especially those defined by Australian, International or American Standards and following discussions with wine producers regarding smoke taint sensory analyses.

It should be noted that a company's acceptance criteria based on data obtained from the sensory methodology are a matter for commercial consideration by that company.

Standard method for sensory analysis of specific attributes.

This is the recommended standard procedure for evaluation of specific sensory attributes in wines such as smoke taint or 'eucalypt' character. Details are provided in Appendix 1 for the example of smoke taint.

1. Assessors

- A panel of 8 to 12 assessors is required. A minimum of eight assessors must be used but 10-12 are recommended. All assessors must have previously demonstrated ability to discriminate and rate the attribute consistently, assessed using screening and qualification tests.

2. General test conditions

- Odour free, quiet, well ventilated, temperature-controlled environment with no distractions.
- Assessors faced away from each other or in booths.
- Samples prepared out of sight of the assessors and in an identical manner.
- Assessment mid-morning or mid-afternoon, no eating or drinking for at least 30 min prior to test.
- Assessment in silence.

3. Sample presentation

- Random presentation order: the order of the samples from one assessor to another should be randomly assigned and preferably balanced to reduce order effects.
- Constant temperature across samples. Samples should be pre-poured early enough to reach a uniform temperature across samples, especially if they came from refrigerated storage, 20-22 °C or 22-24 °C, and constant volume (30 mL).
- Glasses covered with plastic petri dishes or glass watch-glasses.
- Glasses will be coded by random three-digit numbers, and codes should vary for each test.
- Positive and negative controls included. The panel should be aware that controls can be included and expect that not all samples are tainted.
    - Negative control: clean unaffected wine, preferably of the same variety as test samples and made in the same way. A negative control must be included in each set.
    - A positive control ie a wine known to have the attribute should be included on a regular basis.
- Samples expectorated.
- White wines should be tested before reds, with the samples preferably split into similar variety or style.
- Maximum of eight samples per set including controls.
- Samples will be assessed in duplicate. The set of samples will be initially assessed and the test repeated.
- Ensure a break is enforced between samples to minimise carry over effects, which is particularly important in the case of smoke taint assessment. Two minutes between samples and 10 minutes between each set is recommended for smoke taint.
- Encourage water rinsing between samples.
- An unstructured, continuous 15 cm line scale will be used (coded from 0 to 10), with indented anchor points of 'low' and 'high' placed at 10% and 90% respectively (Figure 1). Separate scales

will be used for aroma and flavour for each attribute to be tested as well as an 'other' term. Rating 'overall fruit' is also recommended.

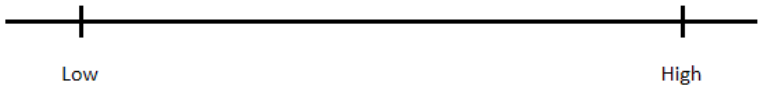


Figure 1. Line scale to be used for rating attributes.

- The performance of assessors will be determined after each test. Panellists' individual agreement with the panel mean, ability to discriminate and repeatability should be assessed through statistical analysis. This data will be used to ensure each judge remains suitable for participation in future tests.

4. Analysis and interpretation of results

- Data will be analysed using Analysis of Variance (ANOVA) with judge, wine and repetition as factors with judges as random effect, and the two-way interactions will also be included in the model.
- If the ANOVA results show that there was no significant difference among the wines for the attribute, the wines in the test set are considered unaffected.
- If the ANOVA results show that the wines significantly differed in the attribute, a specific wine is considered affected if the mean attribute rating is significantly higher than the included clean control as determined from a subsequent Dunnett's means comparison test ($P<0.05$).
    - Alternatively, if consumer acceptance or other background data are available then mean values above a certain agreed value can be reported as affected (e.g. scores greater than 3 tainted, 0-3 acceptable).

Appendix 1. Example of determination of smoke taint using the standard method.

A responsible sensory person (panel leader) is accountable for the panel setup: setting up the program, facilities, training and maintaining panel, analysing and disseminating results to management. The panel leader will identify the venue and the frequency of tests. The panel leader should have excellent organizational capabilities, technical knowledge and knowledge of the sensory methodology, interpretation and reporting of results. Competent communication, writing and decision making skills are also important (International Organization for Standardization 2006a).

1. Assessors

Recruit at least eight qualified assessors, preferably ten. It is recommended at least twice the number of candidates are initially screened to achieve desired numbers.

Inclusion criteria:

a. Sensory acuity to smoke taint compounds

Initial screening test to determine sensory acuity

a. Triangle test (International Organization for Standardization 2004) – two samples of the test material (smoke glycoside spiked sample) and one sample of water, or two samples of water and one sample of the spiked sample are presented to each candidate. Assessor needs to select the 'odd' sample in each set and comment on the main characteristic of the different sample.

- Practical suggestion: Two sets with guaiacol glycosides and two sets with cresol glycosides
- Concentration of glycoside should be at above-threshold level (0.5 mg/L). An inability to detect differences indicates the assessor's lack of suitability for this test. Candidates should preferably have 100% correct responses in a series of repetitive tests (Standards Australia 2014a). A 3 out of 4 correct results might get a pass if the taster has written smoke-related terms in the comments for the correct samples.

b. Assessor availability

c. Limited sample background knowledge: may result in bias when evaluating wines

2. Training

Familiarisation with smoke characteristics

- Aroma identification exercise – present samples of guaiacol (smoky aroma) and cresol (medicinal, band-aid aroma) volatiles spiked in base wine at the following concentrations, which are considered moderate levels of smoke and medicinal sensory characteristics:
    - Clean Control - base wine, relatively neutral bag-in-box wine or similar
    - Guaiacol - 80 µg/L
    - Cresol mix:
        - o-cresol - 70 µg/L
        - p-cresol - 70 µg/L
        - m-cresol - 30 µg/L

Familiarisation with sensory methods

- Materials: Base wine, phenol compounds to spike at the specific dilution, beaker (if not spiking directly into the bottle), wine glasses (International Organization for Standardization 2006b)

- Panellists should be familiar with the scale and verbal anchors. See (Standards Australia 2014a) for training in use of scales – assessors should be familiarised with the concept of the scale by initially rating a set of samples spiked with increasing concentrations of smoke volatiles (four samples).
- Ensure panellists are exposed to spikes of guaiacol and the cresols and familiarised with real smoke tainted samples at low, medium and high levels
- Encourage panellists to think about the presence of smoky, medicinal, ashy, phenolic tarry aroma or flavour, and avoid considering too closely bitterness, drying, harsh palate or 'dirty' aftertaste in the absence of smoky flavour, as they are common reasons for false positives.
- Procedure: present the samples one at a time to assessors and instruct them to smell each sample and make notes on the perceived aromas. The assessor should identify the substance by its chemical name or descriptor (Standards Australia 2014a).
- Remind panellists that not all wines are tainted and there may be clean control wines included in the assessment so scores of zero should be expected in some cases.

Final qualification

- Evaluation of real wines with different levels of smoke taint. It is important to have the assessment of at least one set of six wines in duplicate before the first real sample set. This session is preferably done with real, highly smoke-tainted wine (100% smoke), and diluted with unaffected wine down to 50%, 25%, 12.5%, 6.25%, and 0% smoke. Alternatives include sourcing five smoke affected wines of varying intensity, together with a clean control, or use spikes listed above if smoke tainted wine is not available. Panellist's performance will be assessed for discrimination and agreement with the panel mean, and those assessors who perform poorly will be either subject to further training or excluded from the panel (Standard Australia 2014c).

3. Decision criteria and risk assessment

- An action standard that will determine the sensory conclusion should be specified prior to the test
- Use average scores, decide on action standard for tainted: significantly higher than the included control, or low (e.g. 1.5-2.5), moderate (2.5-5.0) and high (6.0 and above) average scores.
- An individual company's acceptance levels are a matter for commercial consideration, beyond the scope of this procedure, and should be specified with stakeholder agreement.
    - one AWRI consumer test has shown significant decrease in consumer liking for wines with smoke rating of 2.3 or above using this methodology (*AWRI fact sheet: consumer acceptance of smoke-affected wines)*

4. Data collection

Good sensory practices should be applied throughout the test: tasting area odour-free, no noise, 2-minute rest should be emphasised to judges to minimise carry-over effects, no drinking or eating for at least 30 minutes prior to assessment.

- Inform participants of date, time and location of test
- Define sample set – include negative controls and duplicate samples to manage biases.
- Record responses independently (ASTM 2020). Electronic systems will facilitate data collection and data analysis.

Assessor's requirements

- refrain from smoking, eating or drinking (including coffee) for at least 30 minutes before a test (International Organization for Standardization 2012).
- assessors suffering from emotional upsets, colds or other illnesses should be excluded from tests until they recover (Standards Australia 2014b).

- check that assessors are motivated to continue to participate in further tests (Standards Australia 2014b).
- confirm judges are familiar with smoke taint aroma and flavour, and that they have been screened for their ability to perceive flavour from guaiacol glucoside and have been exposed to spikes of guaiacol and the cresols.

Room setup

Equipment

- standard tasting glasses
- black odour-free non-permanent marker pen (whiteboard marker)
- 30 ml graded plastic pourer or shot glass
- water glasses
- spittoons
- iPads/personal devices equipped with data collection software
- paper forms and pens if no software is used

Environment

- constant temperature: 22-24°C
- uniform lighting
- clean area, free from any odours
- quiet atmosphere

Preparation

- Minimise panel interaction with adequate separation of the assessors
- Assign each wine an individual 3-digit code (including repeats). Use computer data acquisition software (*See Appendix 4. Sensory software*) or *University of Oregon link* to generate 3-digit codes or see *Appendix 5. List of random 3-digit numbers.* Write the codes on the bottles to be poured.
- Code rows of glasses with a non-permanent whiteboard marker with the 3-digit sample code for each wine, with enough glasses for the number of judges, and align wine bottles on the bench with the corresponding coded glasses.
- Pour samples, equal volumes (30 mL), and distribute samples to judges.
- Each judge should have a different order, not starting in different positions of the same order. See example of a randomised presentation order below. Advise judges to shuffle samples randomly or the person setting up does this. Distribute devices with data collection software or paper forms and pens to each panellist.

Smoke Panel Session: 22-06-2020

Sample Name	Code S1	Code S2
Semillon Clean Control	734	285
Semillon 7092-002	569	981
Semillon 7094-002	819	637

Sample Name	Code S1	Code S2
Chardonnay Clean Control	213	715
Chardonnay 7093-002	576	809
Chardonnay Smoke Spike	472	126

	Set 1						Set 2					
Tray No.	1	2	3	4	5	6	1	2	3	4	5	6
1	3 -- 819	1 -- 734	2 -- 569	1 -- 285	2 -- 981	3 -- 637	1 -- 213	2 -- 576	3 -- 472	3 -- 126	2 -- 809	1 -- 715
2	1 -- 734	3 -- 819	2 -- 569	2 -- 981	3 -- 637	1 -- 285	3 -- 472	1 -- 213	2 -- 576	2 -- 809	1 -- 715	3 -- 126
3	1 -- 734	2 -- 569	3 -- 819	2 -- 981	1 -- 285	3 -- 637	3 -- 472	2 -- 576	1 -- 213	2 -- 809	3 -- 126	1 -- 715
4	2 -- 569	1 -- 734	3 -- 819	3 -- 637	2 -- 981	1 -- 285	2 -- 576	3 -- 472	1 -- 213	1 -- 715	3 -- 126	2 -- 809
5	3 -- 819	2 -- 569	1 -- 734	3 -- 637	1 -- 285	2 -- 981	1 -- 213	3 -- 472	2 -- 576	1 -- 715	2 -- 809	3 -- 126
6	2 -- 569	3 -- 819	1 -- 734	1 -- 285	3 -- 637	2 -- 981	2 -- 576	1 -- 213	3 -- 472	3 -- 126	1 -- 715	2 -- 809
7	1 -- 734	3 -- 819	2 -- 569	3 -- 637	1 -- 285	2 -- 981	3 -- 472	1 -- 213	2 -- 576	1 -- 715	2 -- 809	3 -- 126
8	3 -- 819	2 -- 569	1 -- 734	2 -- 981	3 -- 637	1 -- 285	1 -- 213	3 -- 472	2 -- 576	2 -- 809	1 -- 715	3 -- 126
9	1 -- 734	2 -- 569	3 -- 819	2 -- 981	1 -- 285	3 -- 637	3 -- 472	2 -- 576	1 -- 213	2 -- 809	3 -- 126	1 -- 715
10	2 -- 569	1 -- 734	3 -- 819	3 -- 637	2 -- 981	1 -- 285	2 -- 576	3 -- 472	1 -- 213	1 -- 715	3 -- 126	2 -- 809
11	3 -- 819	1 -- 734	2 -- 569	1 -- 285	3 -- 637	2 -- 981	1 -- 213	2 -- 576	3 -- 472	3 -- 126	1 -- 715	2 -- 809
12	2 -- 569	3 -- 819	1 -- 734	1 -- 285	2 -- 981	3 -- 637	2 -- 576	1 -- 213	3 -- 472	3 -- 126	2 -- 809	1 -- 715

Figure 2. Example of six samples presented in duplicate to 12 judges in randomised order, over two brackets.

Assessing the wines

- Remind judges there are clean control wines included in the assessment and scores of zero should be expected.
- Judges should concentrate, be honest with themselves, not be too critical and not gravitate to the middle of the scale.
- If using paper forms, remind judges to confirm and record the 3-digit sample code on the evaluation sheet and self-enforce minimum of 2-minute breaks and water rinses between samples (more time might be needed after a heavily affected sample). All assessments should be conducted independently and blind (no knowledge of samples, regions, producers), no discussion during the assessment.
- The scale shown in Figure 1 should be used.
- Close session in data collection software and open results, provide feedback to assessors. If using paper forms collect sheets and confirm all samples were assessed.
- If using paper ballots measure distance on the line scale, enter judge scores, sample identification and 3-digit codes into a spreadsheet, or
- If using data acquisition software, export the results from the software

5. Results and Decisions

- Systematising the standard reporting formats and processes for data handling, recommendations, and action criteria is a very important step. This activity should also include mechanisms for archiving results and tracking both products and panellist performance across time (Lawless and Heymann 2010).
- Results should be analysed by analysis of variance (ANOVA), for the fixed effects of wine and presentation replicate, the random effect of judge, and their two-way interactions, followed by Dunnett's means comparison test to determine whether the samples were rated significantly higher than the control(s).

The panel leader should be prepared for inconsistent pattern in the responses or lack of agreement (E3093-20). Below are some reasons for lack of panel agreement and ways to minimise it:

a. differences in sensitivity among panellists: re-train panellists with spiked samples of smoke glycosides and smoke volatile compounds (guaiacol and cresols) and ensure all assessors are sensitive to the compounds

b. product differences that are too small to be reliably detected: ensure there is a clean, untainted control and try to have a positive control (known tainted sample of same variety/style) in each test. Use as many qualified assessors as possible in each evaluation.

c. lack of consensus in using the smoke taint term: show real smoke tainted samples from different styles of wine, spiked samples with smoke volatiles together with wine spiked with different taints like sulfur compounds that could be confused to smoke taint.

d. presence of off-flavours that could mask smoke taint perception: ensure there is a space to rate *Other* term in the tasting form, with a comment box to specify the off-flavour so this information is captured.

e. false positives caused by bitterness, drying, harsh palate or 'dirty' aftertaste or 'reduced' samples (struck match, burnt toast characters)

f. lack of consensus in the determination of the size of difference, using scale differently: show known tainted samples at low, moderate and high levels to panellists and have the panel agree with their smoke score on the scale

- Cross-reference sensory panel results with corresponding chemical data when available.
- Use email, Word document, and/or Excel to present both sensory and chemical data sets to the responsible party to make a business decision and take appropriate action, depending on the risk tolerance of the business and contractual agreements.

6. Motivation and Panel Performance

- Appreciate and acknowledge the efforts that panellists undertake to attend the sessions, remind panellists regularly of their contributions (E. De Vos, 2010). Provide performance feedback to judges constructively, being encouraging and thankful for their time and effort. Keep individual performance results confidential. Consider small treats for a participation incentive as assessments can be fatiguing and unpleasant.
- Provide information of the panel results to panellists and provide feedback on individual results, and rewards when possible (Standards Australia 2014a).
- Panel performance: monitoring assessor's performance to check they are repeatable, discriminatory, homogeneous and reproducible (Standards Australia 2014a).
- Individual: assess the individual judge's performance by checking negative and positive control sample/repeat scores. Assess the panel precision by checking scores/conclusions on the repeated samples.

  g. Compare individual scores with panel mean

  h. Visual representation of standard deviations

  i. Homogeneity of scoring with respect to the group (overscoring, underscoring)

  j. Individual repeatability, reproducibility

- Panellists should participate on a regular basis to maintain their performance. Weekly participation is desirable, with a minimum of monthly participation. Re-train assessors after long periods of interruption (6 weeks or higher).

Method validation

Test-retest reproducibility

For a set of eight Pinot Noir wines with varied smoke taint, including clean controls, assessed by two separate panels following the above procedure a week apart, the correlation coefficient (r) between the two sets of smoke flavour scores was 0.99 ($P < 0.001$), showing that the test was highly repeatable. An Analysis of Variance determined that there was no significant difference between the two panels in their pattern of rating of the wines (panel x wine interaction $P=0.39$), with both panels having very similar standard error values, indicating the same degree of discrimination.

Discrimination

For six sets of 7-12 varied smoke tainted wines of sets of varieties (Pinot Noir, Chardonnay and Shiraz) assessed using the method, using a minimum of eight assessors, the least significant difference between means for smoke flavour was a maximum of 1.3 units on the 10-point scale, indicating that two means further than this value can be considered to be significantly different in smoke flavour. Clean controls were rated with mean values from 0.2-2.5, median 0.9.

Statistical power

For a set of eight Pinot Noir wines with varied smoke taint, including clean controls, the power of the test to be able to detect a mean difference of at least 1 unit was 1.0, and of 0.5 units was 0.86, showing that the panel of 10 trained assessors, assessing in duplicate, given the error in the test, was able to detect this degree of difference with great confidence.

It was observed that with 7 or fewer assessors there was a loss of power, with wines that were significantly higher in smoke flavour with 10 assessors then not significantly different.

Correlation with wine phenol concentration

For a set of eight Pinot Noir wines with varied smoke taint, the correlation coefficient (r) between smoke flavour scores using this method and smoke-related phenol compounds was greater than 0.91 ($P < 0.001$). For example, correlation with smoke flavour mean scores and the concentration of guaiacol, m-cresol and total glycosides was 0.94, 0.97 and 0.92, respectively.

Appendix 2. Small panel indicative method

If insufficient resources are available to perform the standard sensory procedure, a small panel solution can be considered. Although the data obtained is not as reliable or sensitive as the standard method, it is an improvement on running only informal sensory evaluation or simply relying on the opinion of individuals. Nevertheless, the small panel option should be reconsidered when a high-risk business decision is involved and the data should be treated qualitatively.

- Small panel in this context is a panel with three to seven tasters (ASTM 2020). A panel size of five or more is recommended to be able to detect relatively small differences among wines.
- A 6-point category scale, illustrated below, is recommended and the sum of each smoke taint rating is compared with a pre-decided criteria for tainted wine.

Smoke taint sensory assessment

Date:_____ Judge:_____

Please rate each sample for degree of *smoke taint*

578	☐ no smoke	☐ very slight smoke	☐ slight smoke	☐ moderate smoke	☐ large smoke	☐ very large smoke
396	☐ no smoke	☐ very slight smoke	☐ slight smoke	☐ moderate smoke	☐ large smoke	☐ very large smoke
403	☐ no smoke	☐ very slight smoke	☐ slight smoke	☐ moderate smoke	☐ large smoke	☐ very large smoke

Please rinse with water and wait 2 minutes before assessing the next sample

- The same sensory best practices are recommended in this scenario: samples coded with a 3-digit blinding code, random presentation order, uniform appearance, temperature and volume across samples, ensure a considerable break is enforced between samples to minimise carry over effect, 2 minutes between samples and 10 minutes between each bracket, with plenty of water rinsing between samples. Odour free, well ventilated, temperature-controlled environment with assessors faced away from each other if possible and assessment in silence. White wines should be tested before reds, preferably split into variety or style
- Carefully consider maximum number of samples for the time allocated for the session. For a one-hour session, maximum of 12 wines, divided in two or three brackets are recommended, including repeated samples and clear controls.
- For a small number of samples, a balanced presentation order of the samples should be followed (every possible order occurs an equal number of times).
- At least one sample should be repeated in the set to check panellist repeatability.
- Include positive and negative controls to indicate the rate of false-negatives and false-positives. The panel should be aware of the inclusion of controls and expect that not all samples are tainted.

- negative control: clean unaffected wine, preferably of the same variety as samples in the set and made in the same way).
- positive control: known smoke affected wine. If controls are not available designate one wine in the set to be used as a repeat, i.e. duplicated.

- Action standard to be decided prior to the test.

Data Analysis

Table 1 details the way to process the data and make a conclusion.

Table 1. Smoke taint data evaluation for a 6-point intensity scale - adapted from ASTM (2020), number of assessors for each category.

Smoke taint rating scale	No smoke - 0	Very slight smoke - 1	Slight smoke - 2	Moderate smoke - 3	Large smoke - 4	Very large smoke - 5	Sum	Conclusion
Wine 1	4	2	–	–	–	–	2	Clean
Wine 2	4	2	–	–	–	–	2	Clean
Wine 3	–	–	3	3	–	–	15	Tainted
Wine 4	–	–	3	3	–	–	15	Tainted

- Numbers in the *Smoke taint rating scale* columns are the frequency of tasters assigning the products that value.
- Values in the *Sums* column are calculated as the sum of each smoke rating value multiplied by its frequency. Sum = ∑ (Smoke taint rating x Frequency).
- For example, Wine 1 Sum = (4 x 0) + (2 x 1) = 2
- Conclusion will be based on pre-established decision criteria of a limit Sum value for a sample to be considered Clean. The cut-off number should be defined by management with the agreement of stakeholders.
- Retesting should be considered for inconclusive results.
- Performance of selected assessor should be monitored continuously after each test to ensure the criteria continue to be met (Standards Australia 2014a).
- Data from judges performing poorly can potentially be excluded, without reducing panel size below three. If it is noted that a judge consistently performs poorly, consider offering him/her to help with setting up and analysing results.
- Panellists should be called on a regular basis to maintain their performance. Weekly participation is desirable, with a minimum of a monthly participation. Re-train assessors after long periods of interruption (6 weeks or higher).
- Tests conducted with an AWRI smoke panel for a set of Chardonnay and a set of Pinot Noir wines, comparing the small panel 6-point scale with the standard large panel duplicated line scale method described above, showed a correlation coefficient (r) of 0.79 for Chardonnay ($P<0.05$) and 0.98 for the Pinot Noir wines ($P<0.001$). A higher occurrence of false positives (rating clean control sample as tainted) was observed with the small panel method, plus a considerable loss of sensitivity for less obviously tainted samples. As the panel size decreased, the power of the test also decreased, with correlation coefficient as low as 0.45 for Chardonnay with a panel of 3 assessors. The same conclusion as with the more robust large panel method was established about 60% of the time for a panel of six, 30% for a panel of five and only 13% for a panel of three tasters.

Appendix 3. Overview of sensory procedures for quality control.

Traditional quality control (QC) involves three major requirements: the establishment of specifications, which are the characteristics of the ideal or usual product; a sampling plan appropriate to the product being manufactured or the system being monitored; and the establishment of tolerance limits, which are the levels of variability and/or the ranges that are deemed acceptable (in specification) versus unacceptable (out of specification).

To set the tolerance limits, the liabilities of Type I error, rejecting product that is acceptable and therefore incurring unnecessary cost, must be weighed against the potential for Type II error, letting bad product into the marketplace and offending loyal consumers. This is a management decision that will impact the nature of the tolerance limits that are set (Lawless and Heymann 2010). Often a combination of sensory and instrumental measurements is used to obtain a robust quality measurement, and the attribute acceptable range should reflect consumers' expectations of the product (Bereen 2010)

The success of any program demands strong management commitment. Without management support, a sensory QC program is bound to fail (Lawless and Heymann 2010).

Good sensory practices

The use of good sensory practices ensures a QC sensory test gives an objective outcome that is not based on taster's individual technical experience and familiarity with production, nor based on the highest-ranking person on the panel. Muñoz (1992) and Lawless and Heymann (2010) stressed the need for standardised protocols for sample handling and evaluation and the need for independent judgments, rather than discussion and consensus.

An example comparing two sensory analysis procedures shows how good practices generate reliable results even with a very small (three people) panel as described by Lawless and Heymann (2010). In the first test type, three to five panellists assessed boar taint aroma in a pork meat packing facility and came to a consensus judgment after applying a 6-point scale for odour intensity. The second evaluation procedure was done in the analytical laboratory well ventilated area by three panellists, who were screened for sensitivity to the taint. Means were calculated from a 9-point scale for odour intensity with standardised preparation procedures. The correlation coefficient with instrumental measures of the taint was +0.27 for the first evaluation (not significantly different from zero correlation) and +0.40 in the second (statistically significant). The methodological improvements in the second evaluation explain the increased correlation, including (1) better evaluation location (laboratory versus packing house), (2) screening of judges, (3) constancy of panel members instead of people dropping in and out, (4) averaging scores versus a consensus procedure, and (5) a more standardized sample preparation method. This made a difference in the instrumental-sensory correlations.

It is well established that evaluations should be made in a clean, well-ventilated, temperature-controlled environment, preferably with booths or separators, not on the manufacturing floor. Facilities should be odour free and distraction free (Lawless and Heymann 2010). Product samples should be blind coded and presented in different random orders to each panellist. Serving temperature, volume, and any other details concerning product preparation and the tasting method should be standardised and controlled as far as possible, including air and light quality, limited noise, distraction and space requirements (International Organization for Standardization 2017).

The presentation of blind control samples is necessary to estimate a false alarm rate (false positives), and the introduction of purposely defective samples is useful in estimating the false-negative (miss) rate (Lawless and Heymann 2010). Blind replicates are also recommended to check judge consistency (Muñoz 1992), if all samples cannot be assessed in replicate.

General sensory quality control guidelines (adapted from Lawless and Heymann (2010):

1. Establish standards for optimum quality ('gold standard') target plus ranges of acceptable and unacceptable products
2. Standards should be calibrated by consumer testing if possible. Alternatively, experienced personnel may set standards, but these should be checked against consumer opinion (users of product)
3. Judges must be trained, i.e., familiarized with standards and limits of acceptable variation
4. Unacceptable product standards should include all types of defects and deviations likely to occur from materials, processing, or packaging
5. Judges may be trained to give diagnostic information on defects if standards are available typifying these problems.
6. Ideally, statistically meaningful data should be gathered (ten or more observations per sample)
7. Test procedures should follow rules of good sensory practice—blind testing, proper environment, test controls, random orders
8. Blind presentation of standards within each test should be used to check for judge's accuracy.
9. It is important to include a (blind) gold standard for reference purposes as well
10. Judge reliability may be tested by blind duplicates
11. Panel agreement is necessary. If unacceptable variation or disagreement occurs, re-training is warranted

The sensory panel

Ideally, a panel should operate with at least eight assessors, preferably ten or more (Lawless and Heymann 2010), but methods are available for small panels of three to seven. For any sensory panel, the assessors must have previously demonstrated ability to discriminate and rate the attribute consistently, assessed using screening and qualification tests. The panel should contain enough members to minimise variability in results, however smaller but highly trained and experienced panels can be more reliable than panels with more members but who are less trained and less sensitive (Vos 2010).

It is well established that "variability in sensory response is inherent in any group of people used for testing and is unavoidable; this can arise from inconsistencies within an individual, and through physiological and psychological differences between individuals. However, with training, such a group can show highly consistent … responses." (AS 2542.1.1 2014)

It is unrealistic to expect to have access to a large panel for every evaluation, especially within wine companies, and a compromise can be achieved for small panels of at least three assessors (preferably at least five). Fewer than three people is not recommended as it will not allow a pattern to emerge or a majority to be determined (ASTM 2020). Odd numbers of assessors make it easier to establish the majority. The small number of participants makes it essential to consider all points of view before making conclusions or decisions.

Panellists should be screened and trained to be assured they are qualified for the test, and especially their ability to recognise and rate specific attributes. The same standardised procedures should be used for each test, so results are comparable.

A sensory responsible person should be appointed, who will work as a panel leader and will be responsible for setting up the panel – introducing the program, organising facilities, planning the

training and maintaining the panel, analysing and disseminating the results (Vos 2010). External sensory specialists can help initiate, implement and validate the program.

General guidelines for individual's participation in sensory assessments, from Lawless and Heymann (2010):

1. Be in correct physical and mental condition
2. Know the score card
3. Know the defects and the range of probable intensities
4. Pay attention to the sequence of flavours
5. Rinse, occasionally, as the situation and product type warrant
6. Concentrate. Think about your sensations and block out all other distractions
7. Do not be too critical. Also, do not gravitate to the middle of the scale
8. Do not change your mind. Often the first impression is valuable, especially for aromas
9. Check your scoring after the evaluation. Get feedback on how you are doing
10. Be honest with yourself. In the face of other opinions, "stick to your guns"
11. Practice. Experience and expertise come slowly. Be patient
12. Be professional. Avoid informal lab banter and ego trips
13. Insist on proper experimental controls—watch out for benchtop "experiments"
14. Do not smoke, drink, or eat for at least 30 min before participation
15. Do not wear perfume, cologne, aftershave, etc. Avoid fragranced soaps and hand lotions

Judge selection

A sensory analysis panel is a 'measuring instrument' and therefore the results will depend on its members. The recruitment of panellists willing to participate should be carried out with care, considered as a real investment in both time and capital, with management support from the organization required to be effective (Standards Australia 2014b). Individuals who cannot follow directions or participate in a discussion without emotion should not be included in the evaluation (ASTM 2020). As a general rule, at least two times the number of assessors required in the final panel should be recruited.

Selection of final panellists should be based on interest and motivation; availability; good general health; ability to communicate; knowledge of the product; punctuality, and reliability. While staff can be directed to participate, if they are not well-motivated their performance will not be optimal. It is not essential to include only winemaking or other technical staff if qualification steps are followed: laboratory, packaging or administration staff can be suitable. Screening tests should be conducted aimed at determining sensory acuity (sensitivity to the stimuli under investigation) (Standards Australia 2014b).

Use of action standards

The criteria for acceptance/rejection must be agreed by management or the parties involved (grower and winemaker, winery and customer etc) prior to the test so conclusions from the test can be objectively interpreted. The way to assess whether a sample is acceptable or not will largely depend on the size of panel.

a) Large panel (eight or more panellists): agree on action standard for absent or present e.g. mean score equal to or below 2.5 is acceptable, above 2.5 is tainted; or alternatively low, medium and high mean taint scores. A statistical comparison between the scores of a known clean unaffected wine compared to a test wine is often the most suitable, where if the mean value of a test wine is significantly different to a control the wine is considered affected by the off-flavour.

b) Small panel (three to seven panellists): if the size of panel is too small for statistical analysis (less than eight judges), a decision criteria needs to be set prior to the test. Often a go/no go rule is applied (Meilgaard 2016) e.g. if 50% or more of the panel classify the sample as clear then

the sample is considered accepted. A traffic light system is an alternative approach, for example: less than 40% of the panellists classify the sample as clear then the sample is unacceptable (red), more than 60% is considered acceptable (green), while between 40% and 60% is considered marginal, or inconclusive and requiring retesting (yellow). It is important to set an action standard for inconclusive results. For example, the product can be put on hold until a final decision is made. There is no statistical treatment of responses from a small panel, the intent is to provide a rapid response that is not dependent on a numerical value. If a decision cannot be made, then inconclusive samples could be re-evaluated by a second group of assessors on a later occasion (Stone et al. 2012)

Response scales

The form that will be used to record responses should be simple and focused on the issue at hand to enable efficient collection and collating of responses. Standards Australia (2014b) provides comprehensive general scale requirements in sensory evaluation:

- Scale should be easy to use, discriminating, unbiased and easily understood by the assessors. There may be a choice among several equally good scales
- Rating and scoring are used to evaluate the intensity of one or more attributes, giving an ordinal estimate of the magnitude of attributes, which can be compared to other sets of samples rated with the same scale
- assessors should evaluate the samples independently, one-by-one, in the order that was prescribed to them

- Line scale (continuous, unstructured) vs numerical category scale (discrete) (Standards Australia 2014c).
    - Continuous: unstructured line scales (15 cm), labelled at each end with extreme values of the attribute being assessed. Assessors response is a mark on the scale corresponding to the perceived intensity, which is converted to a number by the analysis. Expresses small differences, but task may be more difficult than a category scale, requiring greater experience of the assessor with the scale and data entry takes longer, unless is acquired using software.
    - Discrete or category scales: 9-point rating scales are more discriminating than those with fewer categories, while response time and repeatability are independent of the number of graduations. A small number of categories such as four or five diminishes discriminatory capacity by increasing ‘end effects’ where assessors are less likely to use the lowest or highest categories. For assessors with little training it is considered that a discrete category scale is easier to use than a continuous scale.
    - A line scale can be used if there is a computerised data collection system, otherwise a category scale is recommended due to its simplicity
    - Irrespective of the scale, the quality of the data will depend on 1. Experience and training of assessors to use the entire scale in an aligned manner, and to associate particular levels of perception with corresponding scale values and 2. How the samples are presented.

Binary responses such as yes/no, pass/fail, in/out or true to target have been suggested for a small group panel as they are simple to setup, train panellists and analyse the results (Rogers 2010). However, they only give answers that the samples are not acceptable, with no information on the degree of difference or intensity of an undesirable attribute. Therefore, information would be missed if a sample presents low perceivable taint and could be potentially remediated.

A target scaling method has been recommended for the measurement of particular attribute(s) that are key to consumers’ liking or disliking of the product (Rogers 2010), providing a quantitative measure of the intensity of the attribute. Samples with known intensities can be used for panel monitoring. Sensory data collection systems are used to make it easier to gather and analyse data. In case of a small panel

(less than ten tasters), simple category scales are recommended, minimising the number of categories (ASTM 2020).

A review of the main differences and utilisation of the binary or scaling methods is presented in the next sections.

Binary (In-Out/Pass-Fail/True-to-Target) method

A trained panel evaluates the samples as being either within specifications (in/pass) or outside (out/fail). Results are reported as the percentage of panellists who classify the product to be 'in' specification, 'clean', or 'true-to-target'(Muñoz 1992).

The major advantage of the in/out procedure is its simplicity. It is especially suited to simple products or those with a few variable attributes. It requires the least amount of panel training and short time needed to perform the test. This is a decision-making tool and should not be used as a rigorous source of product information (Meilgaard 2016), as no quantitative descriptive information about the product is provided. Fixed action standards should be applied to the percentage of 'in' classifications (Lawless and Heymann 2010). The 'in' and 'out' guidelines are provided by management with or without consumer information input.

This method relies heavily on the establishment of sensory specifications and clear examples of products with the boundaries of 'in' or 'out', which is normally not straightforward. Several examples of 'out' and 'in' wine would be required for establishing the action standards.

One disadvantage with this method is a high vulnerability to biases from panellists who often can use other unrelated criteria when assessing samples, because of the pressure in making decisions regarding product disposition. For some assessors there can be a strong inclination to indicate 'in', especially if there are serious ramifications for an 'out' response. Therefore, the program needs careful maintenance to succeed (Muñoz 1992).

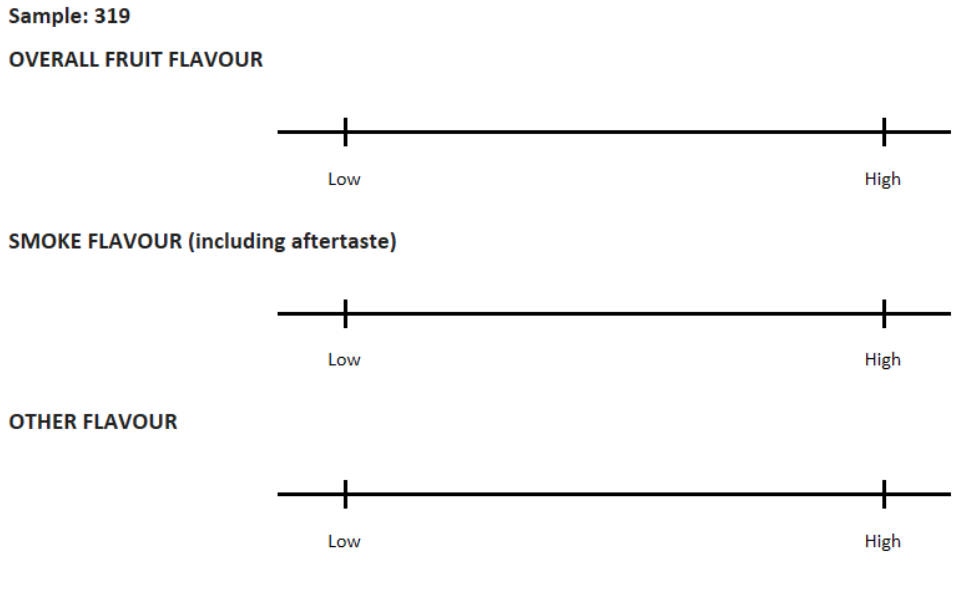
In addition, the method may not provide diagnostic reasons for rejection or failure, so there is a lack of direction to action in fixing the problem. Finally, as the data consist of frequency counts of the number of panellists judging the product out of specification, it is difficult to relate it to other measures such as chemical analyses or consumer preferences (Lawless and Heymann 2010).

Rating of attributes

This section was adapted from Lawless and Heymann (2010). The goal of this type of assessment is to provide intensity ratings for individual key sensory attributes by a trained panel. Intensity rating of single sensory characteristics demands an analytical frame of mind and focused attention on dissecting the sensory experience into its component parts. This method is recommended for situations where the problem areas are well known (intensity of smoke taint, for example) and variability can be reproduced to make up the training set (addition of smoke taint volatiles or glycosides to a wine or water sample).

A

***Smoke**: any type of smoke, including hickory or artificial smoke, phenolic, burnt aroma associated with ashes, ashtray, fire ash, including also medicinal and Band-Aid. Can also include bacon, smoked meat and ashy aftertaste.*



B

Smoke taint sensory assessment

Date:_____ Judge:_____

Sample code:_____ Variety:_____

0: not affected, not noticeable. 1: barely detectable, not sure. 2: perceptible, slight.

5: moderate. 9: high.

	0	1	2	3	4	5	6	7	8	9
Smoke taint	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
Other fault/taint (specify) _____	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Please rinse with water and wait 2 minutes before assessing the next sample

Figure 1 A. A examples of a continuous, unstructured line scale with attribute definition and anchor points used with an electronic system (0 to 10) and B. A paper form discrete structured scale alternative for rating smoke: a 9-point category rating scale with some anchor points.

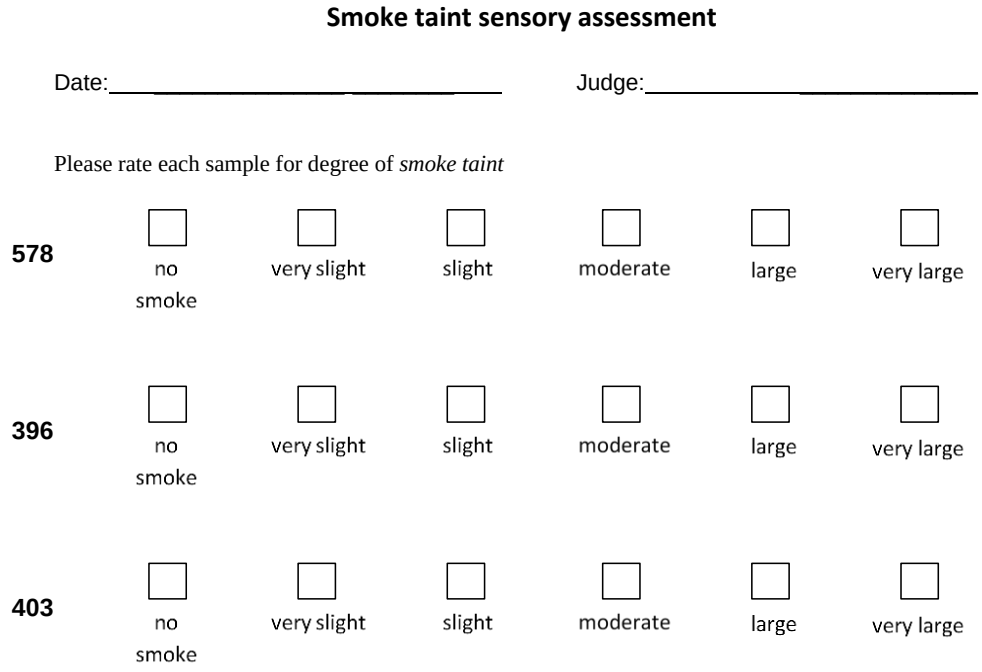


Figure 2. Example of a 6-point category rating scale with anchor points, suitable for small panels with results shown as frequency of counts (ASTM 2020).

As in other techniques, the panel should be calibrated and specification for the accepted profile must be set via consumer testing and/or management input. This will consist of a range of allowable intensity scores for the attribute. Extensive panel training is required, and panellists should be shown reference standards to learn the meaning of the key attributes. Next, they should be shown intensity standards to anchor their quantitative ratings on the intensity scale. For this method they do not have to be shown examples that are labelled as 'in specification' or 'out of specification'.

As with the binary method previously discussed, action standards need to be specified before conducting the evaluation. The decision criteria will vary depending on the size of the panel, which will dictate the ability to perform statistical analysis. In case of a small panel, majority rules normally apply, with for example the number of assessors who gave ratings greater than three, with careful observation of trends or patterns of responses (ASTM 2020). Often, alignment after discussion can be used or re-test is required due to inconclusive results. Exclusion of outlier responses can be considered, but a few low scores (i.e., a minority opinion) may be indicative of an important problem, and re-testing is recommended, especially in situations where there is strong disagreement or high panel variability.

The main advantages of the attribute rating method are: 1. correlation with other measures such as instrumental analysis due to the quantitative nature of the descriptive specification; 2. it presents less of a cognitive burden on assessors, once they have adopted the analytical frame of mind, as they are not required to integrate their various sensory experiences into an overall score, but merely report their intensity perceptions of the key attributes; and 3. the reasons for defects and corrective actions are easier to determine since specific characteristics are rated compared to a pass/fail decision.

The main limitation of the method is that it tends to be laborious in panel training and requires a sufficient number of trained judges. The training regimen is time consuming to conduct, since examples are required for the range of intensities for each sensory attribute in the evaluation. This can also involve substantial technician time in sample preparation.

Conclusion

Adopting proper sensory practices will reduce bias amongst participants, especially those with some prior sample knowledge or a specific agenda (ASTM 2020). Standardised practices are required for sensory evaluation: independent judgements, blind sample coding, limited number of samples evaluated in each session, uniform temperature across samples, proper rinsing and resting intervals, balanced random evaluation order across judges, minimisation of sensory fatigue and carryover, avoidance of sensory and other distractions and use of structured evaluation forms.

If small panels are used, the data should be treated qualitatively, with frequency counts of individual scores reported and considered in action standards.

It is recommended that a test should consist of continuous interval line scale measurements, with a panel of at least eight screened, qualified and trained assessors, preferably 10 or more conducting the test in duplicate. For this type of sensory analysis, statistical analysis is necessary, and data summarized by means and standard errors.

Appendix 4. Sensory software

Listed below are some commercially available software packages that can assist with sensory data collection and data analysis. This list is not exhaustive inclusion on this list does not imply endorsement of any particular vendor.

- Compusense Cloud (Compusense Inc., Canada) – data collection and data analysis https://compusense.com/
- Fizz (Biosystèmes, France) - data collection and analysis https://www.biosystemes.com/en/
- RedJade (RedJade Sensory Solutions, LLC) – data collection and data analysis https://redjade.net/
- Minitab (Minitab Inc., Sydney, NSW) – data analysis https://www.minitab.com/en-us/products/minitab/
- XLSTAT (Addinsoft) - data analysis https://www.xlstat.com/en/
- R with the free SensomineR http://sensominer.free.fr and FactomineR http://factominer.free.fr/ packages – data analysis

Appendix 5. List of random 3-digit numbers

260	149	670	661	698	640	174	444	601	757	420	348
611	616	955	751	597	217	454	366	143	268	447	368
871	222	802	816	388	261	144	219	949	859	437	950
171	435	763	531	317	822	772	940	591	577	265	994
151	117	851	352	128	832	645	766	805	850	321	880
721	558	556	451	312	981	106	799	248	479	262	253
295	982	275	958	132	461	862	452	178	356	433	566
457	934	111	975	209	806	870	373	235	945	939	463
133	524	413	998	325	361	704	899	926	839	340	282
462	449	297	996	718	441	221	504	158	987	823	277
717	752	215	241	655	519	831	848	774	410	360	318
877	345	389	342	727	126	694	168	658	140	588	907
891	436	737	702	573	810	993	161	980	965	584	188
912	300	829	400	786	660	628	273	909	481	157	267
596	349	946	898	505	830	274	391	952	554	257	687
284	642	725	228	878	533	489	104	328	303	863	964
679	777	142	376	364	605	225	175	249	471	735	508

References

Note that Australian, ASTM and ISO Standards are available for purchase online through Standards & Legislation Store APAC – SAI Global Infostore. The Australian Wine Research Institute library can assist with access, and hard copies are available for study in the AWRI library, however due to copyright restrictions, standards cannot be accessed electronically without purchase.

ASTM. 2020. *E3093-20 Standard Guide for Structured Small Group Product Evaluations*. ASTM International, West Conshohocken, PA,. p.

Bereen, C.J.M. 2010. Establishing product sensory specifications. Kilcast, D. (ed.) *Sensory analysis for food and beverage quality control*. Woodhead Publishing Limited.

International Organization for Standardization. 2004. *ISO 4120:2004 Sensory analysis — Methodology — Triangle test*. International Organization for Standardization. 15 p.

International Organization for Standardization. 2006a. *ISO 13300-2 Sensory analysis — General guidance for the staff of a sensory evaluation laboratory — Part 2: Recruitment and training of panel leaders*. International Organization for Standardization. 15 p.

International Organization for Standardization. 2006b. *ISO 5496:2006 Sensory analysis — Methodology — Initiation and training of assessors in the detection and recognition of odours*. International Organization for Standardization. 16 p.

International Organization for Standardization. 2012. *ISO 8586:2012 Sensory analysis — General guidelines for the selection, training and monitoring of selected assessors and expert sensory assessors*. International Organization for Standardization. 28 p.

International Organization for Standardization. 2017. *ISO 8589:2007 Sensory analysis — General guidance for the design of test rooms*. International Organization for Standardization. 16 p.

Lawless, H.T., Heymann, H. 2010. *Sensory Evaluation of Food: Principles and Practices.* 2nd ed.: Springer.

Meilgaard, M.C., Gail & Carr, Bernard. 2016. *Sensory Evaluation Techniques* CRC Press.

Muñoz, A.M., Civille, Gail V. Carr, B. Thomas 1992. "In/Out" Method. Muñoz, A.M. (ed.) *Sensory Evaluation in Quality Control*. Springer, Boston, MA: pp. 140-167.

Rogers, L.L. 2010. Using Sensory Techniques for shelf-life assessment. Kilcast, D. (ed.) *Sensory analysis for food and beverage quality control*. Woodhead Publishing Limited: pp. 143-155.

Standards Australia. 2014a. *AS 2542.1.3:2014 Sensory analysis, Part 1.3: Methodology - Selection, training and monitoring of selected assessors and expert sensory assessors*. Standards Australia. p.

Standards Australia. 2014b. *AS 2542.1.1:2014 Sensory analysis Part 1.1: Methodology - General guidance*. Standards Australia. 20 p.

Standards Australia. 2014c. *AS 2542.2.3:2014 Sensory analysis, Part 2.3: Methodology - Guidelines for the use of quantitative response scales (rating)*. Standards Australia. p.

Stone, H., Bleibaum, R., Thomas, H. 2012. *Sensory Evaluation Practices.* Fourth Edition ed.: Elsevier.

Vos, E.D. 2010. Selection and management of staff for sensory quality control. Kilcast, D. (ed.) *Sensory analysis for food and beverage quality control*. Woodhead Publishing Limited: pp. 17-35.