

Use of feedback calibration to reduce the training time for wine panels

Chris J. Findlay^{a,*}, John C. Castura^a, Pascal Schlich^b, Isabelle Lesschaeve^{c,d}

^a Compusense Inc., 679 Southgate Drive, Guelph, Ont., Canada N1G 4S2

^b Centre Européen des Sciences du Goût (CESG), 15 Rue Hugues Picardet 21000, Dijon, France

^c Cool Climate Oenology and Viticulture Institute, Brock University, St. Catharines, Ont., Canada L2S 3A1

^d Inno Vinum, SDM-RPO, P.O. Box 25009, St. Catharines, Ont., Canada L2T 4C4

* Corresponding author. Tel.: +1 519 836 9993; fax +1 519 836 9898.

E-mail address: cfindlay@compusense.com (C.J. Findlay).

Originally published in *Food Quality and Preference*, 17(3-4), 266-276.

Received 25 September 2004; received in revised form 23 May 2005; accepted 5 July 2005.

Available online 18 August 2005.

Abstract

The performance of descriptive panels is typically determined by post-hoc data analysis. Poor panel performance is measured after the fact and often arrives too late to help the panel leader during training sessions. The feedback calibration method (FCM) optimizes proficiency by ensuring efficient panel training. A previously trained panel (Panel T) and an untrained panel (Panel U) developed and refined their own training targets using FCM before evaluating 20 white wines in triplicate. Permutation tests of the RV coefficient were used to compare the panels in terms of the underlying sensory space. The results of the panels were similar, both Panel T and U were superior to a proficient conventionally trained red wine panel (Panel D). Panel U performed similarly to Panel T on proportion of attributes discriminated and disagreement using a two-way mixed-model analysis of variance (ANOVA) and on multivariate discrimination evaluated by a MANOVA with the same mixed model. Evaluation means for product*attribute fell within the training range targets in 59% of the cases for Panel T and 69% for Panel U, providing an indication of the panels' abilities to hit the training targets. Panel U was shown to be proficient in discriminating a full range of wine attributes ($p = 0.05$) after only nine formal training sessions (22.5 h), a reduction in training time of 49%.

Keywords: Feedback; Calibration; Descriptive; Training; White wine

1. Introduction

A descriptive sensory panel can provide information about the sensory properties of consumer products that is far richer than can be provided by instrumental devices alone, but significant training is required before the panel becomes a reliable sensory instrument. It has been demonstrated that a trained panel can better differentiate products than an untrained panel, and the assessors on a trained panel show greater agreement than assessors on an untrained panel (Wolters & Allchurch, 1994). Product-specific training appears to play a much greater role than sensory experience on unrelated products in obtaining reproducible results. This is supported by the observation that previously untrained panelists who undergo a similar training process to those on a trained panel can be added to an existing panel (Chambers & Smith, 1993). Context effects, which are biases introduced when an assessor perceives the same stimulus differently due to a change in the frame of reference, can be decreased by calibrating assessors on a descriptive panel, but calibration of panelists remains poorly understood (Lawless & Heymann, 1998).

Although there are many ways to communicate performance to panelists, typically a trained panel receives feedback from the panel leader in a group setting after each training session (Meilgaard, Civille, & Carr, 1999). These debriefings provide psychological rewards (Meilgaard *et al.*, 1999),

reinforce training, and increase motivation (Lyon, 2002). Feedback often takes the form of a summary report provided to each panelist at the conclusion of a testing session (Lyon, 2002). Where panelists have evaluated the same product more than once, this report can show means and standard deviations on an attribute-by-attribute basis for both the individual panelist and the overall panel (Meilgaard *et al.*, 1999). Kuesten, McLellan, and Altman (1994a, 1994b) studied the effect of providing panelists with exactly this kind of feedback: panelists were provided with a graph that showed the individual panelist's performance relative to the panel averages and standard deviations for samples evaluated during the session. This experiment looked at two basic tastes (sweet and sour) at various concentrations, and did not conclusively demonstrate whether or not feedback might be capable of reducing variance. Overall, computerized graphical feedback was found to be effective in reducing context effects experienced by panelists responding to line scale questions.

Training times for descriptive panels vary according to the descriptive analysis method used, type of panel, complexity of the product category, number of sensory dimensions in the product, and level of training desired (Muñoz, 2003). A panel leader might expect training to involve 10 h, for a relatively simple product (Stone & Sidel, 1985), to 120 h or more for a more complex product (Meilgaard *et al.*, 1999). Panel-to-panel variability and the training style of panel leaders can influence the outcomes of descriptive analysis. Improvements in the effectiveness of training, whether through time savings or fewer resources invested in the panel, has practical implications for organizations that conduct or intend to conduct descriptive analysis.

There has been a great deal of speculation and research into the role that time plays on the effectiveness of feedback on performance outside the field of sensory science (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Kulhavy & Wager, 1993; Kluger & DeNisi, 1996). In a previous study, two inexperienced panels were calibrated to attribute training targets for 20 commercial red wines established by an experienced determination panel (Findlay *et al.*, submitted). Following a common training period, a group of 16 panelists were divided by lottery. One panel (Panel C) only received numerical feedback and discussion in a group setting. The other panel (Panel E) received immediate graphic feedback during the normal flow of the computerized ballot, but received no further instruction from the panel leader. The panel that only received immediate computerized graphic feedback produced similar results to the more conventionally trained panel, indicating that FCM was a successful approach for training a descriptive panel.

The current experiment was undertaken to further explore the benefits of FCM and to determine its suitability for use in routine sensory analysis training. This study investigated the ability of a previously trained group of panelists and a panel comprised of completely untrained people, to develop and refine their own training targets. We also wanted to investigate the effect of combining FCM with the typical best practices utilized for descriptive panel training. Of particular interest was whether this refinement could reduce the time required to obtain meaningful results from a descriptive panel. In all, the performance of five panels were compared and contrasted to draw conclusions about the suitability of FCM (Table 1).

2. Materials and methods

2.1. Selection and storage of products

White wine is a real-world complex product, making it ideally suited for use in testing the panel training methodology. Vincor International (Mississauga, Ont., Canada) donated 16 products for the study. The Liquor Control Board of Ontario (Toronto, Ont., Canada) provided an additional four products. The 20 commercial white wines encompassed of the following varieties: Chardonnay (4), Riesling (4), Sauvignon Blanc (2), Pinot Grigio (1), Gewürztraminer (1), Sylvaner (1), Vidal (1), Sauvignon-Chardonnay (1), Semillon-Chardonnay (1), as well as blended white wines (4). Wines were stored upside down in their cases in a climate controlled dark room maintained at 20 °C ±1 °C.

2.2. Trained panelists selection, training, and evaluation

Twelve trained red wine panelists were invited to participate in the white wine panel (Panel T) and were paid for their involvement. The panel used the Wine Aroma Wheel (Noble *et al.*, 1984; Noble *et al.*, 1987) and was provided with natural and physical standards to develop their own white wine lexicon and training targets over 5 days of group training sessions of 2.5 h each. They were introduced to the ballot and scaling and given exercises in the booths to develop familiarity with the process. To maintain an accurate record and permit confirmation of the techniques used, all training and debriefing sessions were videotaped. Initial training targets were established by panel consensus in group sessions and through training sessions in booths. These initial values were used in the first FCM session. They refined their training targets through group discussion led by the panel trainer on an iterative basis over the next three training sessions. The target value was expanded around the discrete value by using a range based on the 90% confidence intervals of the previous session's results. This provides a visual zone (Fig. 1) in which any response would be considered to hit the target. Further discussion on the methods to select a range and express the accuracy value of a panelist's response may be found in Castura, Findlay, and Lesschaeve (2005). Each panelist received this feedback on 42, 105, and 262 occasions (an occasion is each time that an attribute is evaluated for an individual product) based on twice the 90% confidence intervals of previous assessments, and 184 occasions based on the 90% confidence intervals, in the final session. The final ballot consisted of 110 attributes evaluated using a structured line scale marked at a quarter, half and three quarters of the line with extreme anchors of "None" on the left, measured as zero, and "Intense" on the right measured as 100 (Fig. 1). Four sensory modalities were used (Table 2): aroma before stirring the glass, aroma after stirring the glass, flavor, and taste/mouthfeel. To alleviate potential dumping effects, the panel was able to indicate the presence of additional descriptors noted during training from an on-screen checklist, as well as to provide comments to indicate the presence of additional attributes. A time delay of 90 s between samples was applied, with a 5-min delay between samples 3 and 4. After nine days of training (20.5 h), Panel T evaluated 20 wines following a complete block design, partially presented over three days. Data were collected for three replicates over a total of nine sessions. Because of the consumption of alcohol, the panelists were required to remain in the training room for at least 20 min at the end of the training session. Food was provided, and panelists were afforded the opportunity to socialize or read magazines.

2.3. Untrained panelists screening, selection, training, and evaluation

Fifty potential panelists were recruited according to the following criteria: non-smoker, available during testing times, drink white wine at least once per month, no previous experience in sensory analysis, no health problems that might interfere with testing, and no specialized knowledge of wines. A two-stage screening procedure eliminated respondents with obvious sensory deficiencies, such as the inability to identify basic tastes, and selected 12 panelists that were able to perform descriptive analysis tasks. The untrained panel (Panel U) also used the Wine Aroma Wheel and was provided with natural and physical standards to develop their own white wine lexicon and training targets over five days of training sessions of 2.5 h each. By comparison, Panel C and Panel E from a previous study (Findlay *et al.*, submitted) were comprised of untrained panelists that had undergone a screening procedure similar to the one used to establish Panel U in this project. For later comparison, Panel C underwent a 40h training regimen that consisted of twenty 2-h sessions. The U panel refined its own attributes and training targets on an iterative basis over four training sessions, during which the panel received feedback on 99, 114, 116, and 100 attribute* products, based on the 90% confidence intervals of previous assessments, which represented a further refinement of the methodology. In summary, they received a total of 22.5 h training over nine sessions. Their final ballot consisted of 76 line scale attributes and they went on to evaluate products exactly as described above for Panel T.

2.4. Training feedback methodologies

Each line scale question screen contained up to five attributes. Immediate monadic feedback was considered, but was felt to be disruptive to the flow of the ballot. Five attributes were chosen as the maximum for feedback on a single screen. Training attributes were selected in each sensory modality to provide a full range of response and over examples that could be generalized to provide guidance for scaling of all attributes. Feedback, when provided, appeared in the form of ellipses on the line scale that indicated the training target range. This occurred immediately after a panelist had evaluated all line scale attributes on the screen (Fig. 1). The panelist's responses remained visible when feedback was shown. Panelists had the opportunity to re-evaluate the sample in the presence of feedback, but not to change their score. This provided immediate individual guidance that permitted discrete self-correction. The panel leader had the opportunity to provide additional feedback and commentary to the panelists in debriefing sessions. Any dispute over a perceived difference between attribute scale values was resolved in the group session. This led to the refinement of target values.

2.5. Sample presentation for evaluation

Twenty samples were presented to panelists using a partial presentation design developed using Design Express 1.5 (Wakeling, 2003); panelists received seven samples on day 1, seven samples on day 2, and six samples on day 3. Wine bottles were placed in the laboratory refrigerator (True Manufacturing Company Inc., O'Fallon, MO, USA) on Friday and kept at 3 °C for the duration of the weekend. On day 1, bottles were removed from the refrigerator 1 h prior to serving. Wines were uncorked 30 min prior to serving and verified free of cork taint by the panel leader. Thirty milliliter samples were poured into clear IMAO wine glasses labeled with three-digit blinding codes. To allow wine samples to be served from the same bottle for the duration of the replicate, bottles were sparged with nitrogen gas, the corks re-inserted, and the bottles replaced in the refrigerator. This technique had been used previously and ANOVA conducted for replication effect showed no significance over sessions. Wine temperature at serving time was 20 °C ± 1 °C. Samples were presented monadically to panelists via serving pass-throughs. Red lights were used to mask wine color in the sensory booths. Between samples, panelists were instructed to rinse their mouths with room temperature distilled water. Unsalted soda crackers, as palate cleansers, were also available to panelists, as was hot water during aroma attributes to allow panelists to humidify their nostrils. The evaluation process was repeated on day 2 and day 3 such that each panelist was presented each of the 20 wines. This entire procedure was replicated three times.

2.6. Data collection and analysis

Data obtained from Panel T and Panel U were collected and analyzed using an enhanced version of Compusense *five* (Compusense Inc., Guelph, Ont., Canada), which was programmed to provide feedback to panelists during the test. During training, frequency counts of panelists who were out of the training range were analyzed using this software, as well as Microsoft Excel (Release 10.6501.6626 SP3, Microsoft Corporation, Redmond, WA). The descriptive analysis data on the 20 white wines evaluated by both panels was collected using Compusense *five* (Compusense Inc., Guelph, Ont., Canada). Two-way mixed-model ANOVA with interaction treating panelists as random effects and wines as fixed effects were conducted using both SAS (Release V8.2, 2001, SAS Institute, Inc., Cary, NC, USA) and SPSS (Release 9.0.1, 1999, SPSS Inc., Chicago, IL, USA), and the results were used as a panel diagnostic. Discrimination was indicated by a significant product effect at $p = 0.05$, and disagreement was indicated by a product*subject effect at $p = 0.05$. Using univariate two-way mixed model ANOVA to select significant attributes at $p = 0.05$, panel average responses over panelist and replicate were used to create data sets for each panel for each of the following sensory modalities: aroma before stirring the wine glass, aroma after stirring the wine glass, flavor, and taste/mouthfeel. Principal component analysis was performed on the covariance matrix (cov-PCA) and biplot projections were used to interpret the resultant sensory space.

It is possible to calculate the degree of similarity between two datasets or configurations by using the regression vector (RV) coefficient. RV can be understood as a correlation coefficient in a multidimensional space ($0 < RV < 1$) the closer to 1, the more similar the configurations or the two sensory spaces. By conducting a large number of random and independent permutations, it is possible to draw the distribution of the RV calculated this way. Using a normal distribution (N), a Z can be calculated and a non-parametric test statistic created. Kazi-Aoual, Hitier, Sabatier, and Lebreton (1995) gave the analytical expression of the mean and variance of the N-RV coefficients calculated by permutation of n lines of one of the datasets to be compared to the other with RV. A normalized deviation between the observed RV and the calculated RV, called the NRV, can be computed. If NRV is >2 , similarity between the two tables can be considered as higher than what would be calculated by chance. Data from Panel T and Panel U can be used to validate one another's sensory spaces: the NRV based on permutation tests was calculated to determine whether the RV calculated was different than an RV generated by chance (Schlich, 1996). An NRV larger than 2 indicates that two configurations are significantly ($p = 0.05$) more similar than when product labels are permuted within one configuration. An NRV between 1.5 and 2 indicates that two configurations are slightly similar, and an NRV between 1 and 1.5 indicates almost no similarity. Covariance PCA biplot projections, RV and NRV calculations were conducted in SAS.

Reduction in training time can be measured in hours, and proficiency can be assessed by examining univariate and multivariate discrimination and disagreement. The target that was established to declare the method a success was a minimum of 25% reduction in training time to yield a descriptive sensory panel with a similar or higher level of proficiency ($\alpha = 0.10$). A panel's ability to develop and refine its own training targets using FCM can be assessed by measuring both univariate and multivariate discrimination and disagreement to determine whether or not these were consistent with those of a more conventionally trained panel. The percentage of evaluation means for product *attribute within the panel's own training range targets can provide an indication of the panel's ability to establish and maintain consistent and accurate training targets.

3. Results and discussion

3.1. Comparing panel T to panel U in terms of the underlying sensory space

Panel-selected attributes were established for each panel for each of the following sensory modalities: aroma before stirring the glass, aroma after stirring the glass, flavor, and taste/mouthfeel. Each data set was submitted to Covariance PCA and the RV coefficient used to compare the eight wine configurations. NRV (Table 4) was calculated to determine whether the RV was different than what might be expected by chance. Results indicate that Panel T and Panel U perceived the 20 wines slightly similarly for aroma before stirring (Fig. 2), and negligible similarity for aroma after stirring (Fig. 3); however, Panel T and Panel U perceived the 20 wines very similarly for flavor (Fig. 4) and taste/mouthfeel (Fig. 5). The lower NRV for the aroma modalities is due to the fact that the T panel duplicated attributes in the three modalities, whereas the U panel evaluated fewer attributes with less redundancy. Wine configurations among the sensory modalities in Panel T were more similar than the wine configurations among the sensory modalities in Panel U. This could indicate that Panel U is more efficient in the sense that its different sensory modalities are less correlated.

As indicated by product p-values for Panel T (Table 2) and Panel U (Table 3), performance was better for flavor and taste/mouthfeel attributes than attributes for aroma, both before and after stirring the glass. This is consistent with expectations: evaluation of a complex product such as wine is more difficult by aroma (Aubry, Etievant, Sauvageot, & Issanchou, 1999).

Multivariate ANOVA were obtained for panel-selected attributes for the univariate ANOVAs. The F-approximation of the Hotteling–Lawley trace and the number of significant canonical variates for each sensory modality provide an insight into both panel proficiency and the perceived complexity of the products (Table 5). Results show the complexities and panel proficiencies are similar,

although Panel U's data suggests the panel is better discriminating and attuned to more complexity in taste/mouthfeel.

3.2. Improving on existing training methods for descriptive sensory panels

The results from Panel U were utilized to determine whether an untrained descriptive panel, using the FCM, produced the same or better results in less time than one trained using more conventional feedback only or immediate feedback only. Panel U was submitted to a 22.5 h training regimen that consisted of nine 2.5 h sessions. In the three evaluation replicates, Panel U discriminated 40 of 76 (53%) attributes and its panelists disagreed on 9 of 76 attributes (12%), both calculated using a two-way ANOVA. Not only was Panel U able to discriminate a higher percentage of attributes after a training period that was approximately half the duration of the training time allotted to similar panels in a previous study (Findlay *et al.*, submitted), but it was charged with a more challenging task: to develop a lexicon based on the Wine Aroma Wheel, then create and refine its own training targets. Using the methodology described here, there was a 48.75% reduction in time required to calibrate the descriptive sensory panel to a similar or better proficiency as judged by two-way ANOVA ($p = 0.05$).

3.3. Developing and refining targets using feedback calibration

A key aspect of this study was to determine if a panel is able to develop and refine its own training targets using FCM. Discrimination and disagreement are indicators of the success of the training methodology, and are reported for number and percentage of attributes used in the evaluation. Discrimination cannot be considered in isolation because it may not be reasonable to assume that all products can be discriminated using particular attributes: intensities may fall within a single just-noticeable-difference interval across the product category, or across the subset of products that were evaluated. In this case, panel agreement suggests a well-calibrated panel, whereas disagreement would call into question the panel's proficiency.

Both the Panel T and Panel U used the Wine Aroma Wheel to refine a lexicon. Training targets were generated by the panel and refined on an iterative basis. The targets were set broadly at first, with examples of low, medium and high intensities. These targets were refined on the basis of actual measurement and adjustment of values following each training session. In some cases, the attribute showed very little variation over the products being assessed and was not used in training.

Panel T discriminated 55% attributes (Table 6), and disagreed on only 4%. Panel U discriminated 53% and disagreed on 12%. In a previous study of red wine, Panel D generated 130 attributes and evaluated them to produce sensory profiles of 20 commercial wines in a more conventional fashion. This panel was judged to be proficient. Its results were used as a benchmark: Panel D discriminated 24% and showed disagreement on 33% of attributes. Comparisons between Panel D from the previous study and Panels T and U are reasonable: both product categories were complex, and although the common attributes were perceived against different backgrounds, the physical and natural examples used to assist panelists in identifying sensory attributes were identical in the case of 22 attributes. These attributes were common for all panels. Panel T discriminated 68% with no disagreement. Panel U discriminated 55% with disagreement on two attributes. Panel D discriminated 50% with disagreement on four attributes. These findings support the results presented when all attributes are considered. Panel T and Panel U were superior to Panel D, indicating that it may not only be possible but advantageous for a panel to develop and refine training targets using FCM. Panel T was slightly better than Panel U. It appears, however, that FCM is effective whether or not the panelists have previous exposure to descriptive analysis.

3.4. Precision and accuracy of training targets

During the last 4 of 9 training sessions Panel T received 593 feedback events, which represented 370 of the 2200 (27%) product*attribute possibilities. During the last 4 of 9 training sessions Panel U

received 459 feedback events, which represented 415 of the 1520 (30%) product*attribute possibilities. Training targets for Panel T were based on twice the 90% confidence interval for all but the last training session; training targets for Panel U were based on the 90% confidence interval for all sessions. Frequency counts of the number of times a panel's evaluation mean across reps fell within the training target provide a measure of the accuracy of training targets, and can be expressed as a percentage of the total number of feedback events (Table 7). When the evaluation mean missed the training target, it was below the range in 98% of cases for Panel T and 93% of cases for Panel U. This indicates that both panels showed a tendency to score below the target value rather than above the target value. It is possible that the tendency to underestimate comes from a desire to err on the side of caution, rather than overestimate the value. Results broadly indicate that both panels were able to create and refine training targets. Evaluation means fell within the training range in 59% of the cases for Panel T and 69% for Panel U, providing an indication of the accuracy of the training targets. On each training session, new target attributes were introduced and during the course of presentation of six or seven samples within a session learning must be taking place. Consequently, it is difficult to gauge the overall effect. Intra-session improvement will be measured in future research. Previously untrained panelists are very receptive to calibrating themselves to targets presented by FCM. It is also worth noting that the previously trained T panel required some "relearning" of scale usage that resulted from their prior experience on other product-specific panels. In the absence of previous experience, the U panel did not require any retraining.

4. Conclusions

Results indicate that the Feedback Calibration Method was effective in training two panels to evaluate 20 white wines, regardless of the panels' level of previous experience with descriptive analysis. FCM, which is undergoing continued refinement, was observed to reduce training time required to train a white wine panel by about one half, when compared to two red wine panels that followed a similar procedure for recruitment and training. Based upon the success of this method with complex products, such as red and white wine, it is likely that all descriptive panels could benefit from an appropriate implementation of FCM during panel training.

One of the major advantages of the method is individual training and self-correcting of panelists. With appropriate calibration standards for attributes, FCM can be used to provide stable descriptive results between panels and over time. Proficiency is addressed during the training rather than post-hoc. A disadvantage might be that proper execution of this method requires careful planning of feedback, since the value of the feedback is dependent on the truth of the value being used. If the values are incorrect, this may lead to considerable confusion on the part of panelists. However, in the hands of competent panel leaders, the method can only enhance panel training.

Further research needs to be done to better understand the levels of discrimination that are possible for any attribute within the context of a specific product category. Additional tools could be developed to provide greater diagnostic information for the panel leader to help in the refinement of attribute targets.

Acknowledgements

The authors are grateful for the funding provided to Project #515979 and Project #474766 by the National Research Council of Canada—Industrial Research Assistance Program (NRC-IRAP). The authors are also grateful to Vincor International Limited and the Liquor Control Board of Ontario for providing the commercial wines that were used in this study. The authors thank Amanda Bartel and Karen Phipps, who were responsible for panel recruitment, screening and training. The authors also thank the sensory panelists whose commitment made this study possible.

References

- Aubry, V., Etievant, P., Sauvageot, F., & Issanchou, S. (1999). Sensory analysis of Burgundy pinot noir wines: a comparison of orthonasal and retronasal profiling. *Journal of Sensory Studies, 14*, 97–117.
- Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213–238.
- Castura, J. C., Findlay, C. J., & Lesschaeve, I. (2005). Monitoring calibration of descriptive sensory panels using distance from target measurements. *Food Quality & Preference, 16*, 682–690.
- Chambers, E., & Smith, E. A. (1993). Effects of testing experience on performance of trained sensory panelists. *Journal of Sensory Studies, 8*, 155–166.
- Findlay C. J., Castura, J. C., & Lesschaeve, I. (submitted). Feedback calibration: a training method for descriptive panels. *Food Quality & Preference*.
- Lawless, H., & Heymann, H. (1998). *Sensory evaluation of food, principles and practices* (1999). Gaithersburg, MD: Aspen Publishers, Inc. (a Chapman & Hall Food Science Book).
- Lyon, D. H. (2002). *Guidelines for the selection and training of assessors for descriptive sensory analysis. CCFRA Guideline No. 37*. Chipping Campden, Gloucestershire, UK: Campden & Chorleywood Food Research Group.
- Kazi-Aoual, F., Hitier, S., Sabatier, R., & Lebreton, J. (1995). Refined approximations to permutation tests for multivariate inference. *Computational Statistics & Data Analysis, 20*, 643–656.
- Kluger, A. N., & DeNisi, A. (1996). The effect of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 2*, 254–284.
- Kuesten, C. L., McLellan, M. R., & Altman, N. (1994a). Computerized panel training: effects of using graphic feedback on line scale usage. *Journal of Sensory Studies, 9*, 413–444.
- Kuesten, C. L., McLellan, M. R., & Altman, N. (1994b). Influence of computerized panel training on contextual effects. *Journal of Sensory Studies, 9*, 401–412.
- Kulhavy, R. W., & Wager, W. (1993). Feedback in programmed instruction: Historical context and implications for practice. In J. V. Dempsey & G. C. Sales (Eds.), *Interactive instruction and feedback* (pp. 3–20). Englewood, NJ: Educational Technology Publications.
- Meilgaard, M., Civille, G. V., & Carr, B. T. (1999). Selection and training of panel members. In *Sensory evaluation techniques* (3rd ed., pp. 133–160). Ann Arbor: CRC Press.
- Muñoz, A. (2003). Training time in descriptive analysis. In H. R. Moskowitz, A. M. Muñoz, & M. C. Gacula (Eds.), *Viewpoints and controversies in sensory science and consumer product testing* (pp. 351–356). Trumbull, NJ: Food & Nutrition Press, Inc.
- Noble, A. C., Arnold, R. A., Buechsenstein, J., Leach, E. J., Schmidt, J. O., & Stern, P. M. (1987). Modification of standardized system of wine aroma terminology. *American Journal of Enology and Viticulture, 38*, 143–146.
- Noble, A. C., Arnold, R. A., Masuda, B. M., Pecore, S. D., Schmidt, J. O., & Stern, P. M. (1984). Progress towards a standardized system of wine aroma terminology. *American Journal of Enology and Viticulture, 35*, 107–109.
- Schlich, P. (1996). Defining and validating assessor compromises about product distances and attributes correlations. In T. Naes & E. Risvik (Eds.), *Multivariate Analysis of Data in Sensory Science* (pp. 259–306). New York: Elsevier Science B.V.
- Stone, H., & Sidel, J. L. (1985). *Sensory evaluation practices*. New York: John Wiley & Sons.
- Wakeling, I. (2003). Design Express 1.5. United Kingdom.
- Wolters, C. J., & Allchurch, E. M. (1994). Effect of training procedures on the performance of descriptive panel. *Food Quality & Preference, 5*, 203–214.

Table 1
 Descriptions of the five panels whose results were compared and contrasted to draw conclusions about the effect of FCM

Panel	Year	Number of panelists	Product evaluated	Description
T	2003	12	20 white wines	White wine panel composed of trained panelists
U	2003	11	20 white wines	White wine panel composed of untrained panelists
D	2002	11	20 red wines	Red wine determination panel composed of trained panelists
C	2002	8	20 red wines	Red wine control panel composed of untrained panelists
E	2002	8	20 red wines	Red wine experimental panel composed of untrained panelists

Table 2
 Final lexicon of the trained panel (Panel T) showing p_{wine} from 2-way ANOVA

Attribute	Aroma Before	Aroma After	Flavor	Taste/ Mouthfeel
Apple ^a	0.2900	0.1268	0.0023	- ^d
Green Apple	0.0356	0.2734	0.0060	-
Banana ^a	0.0932	0.0172	0.1660	-
Grape	0.3061	0.3535	0.0668	-
Peach ^a	0.1005	0.0066	0.0004	-
Pineapple ^a	0.0006	0.0276	0.0016	-
Other Tropical Fruit ^b	0.0001	0.0200	0.0049	-
Melon ^a	0.2085	0.1932	0.0011	-
Pear	0.0541	0.0276	0.0000	-
Lemon Zest	-	0.2509	0.1657	-
Rose ^a	0.0805	0.7085	0.7632	-
Elderflower	0.0001	0.0007	0.0000	-
Honey ^a	0.0691	0.0165	0.0015	-
Butterscotch ^a	0.6294	-	-	-
Brown Sugar	0.0592	0.0668	-	-
Vanilla ^a	0.0018	0.0004	0.0016	-
Alcohol ^a	0.0320	0.0325	0.0000	-
Pungent (irritation from alcohol) ^c	0.0042	0.0061	0.0000	-
Nail Polish Remover	0.4528	0.3172	0.0044	-
Solvent	-	0.0139	-	-
Asparagus ^a	0.0003	0.0036	-	-
Black Pepper ^a	0.0779	0.0353	-	-
Cinnamon	0.0504	0.0016	-	-
Clove ^a	0.7267	0.1846	-	-
Butter ^a	0.1739	0.3329	0.1685	-
Earthy ^b	0.2578	0.0288	0.0000	-
Horsy/Leather ^a	0.0107	0.1867	0.0000	-
Sulphur (burnt matches) ^a	-	0.0001	-	-
Sulphur (cooked vegetables)/ Sauerkraut ^b	0.0000	0.0035	0.0000	-
Turpentine/ Terpenes/ Pine resin	0.0613	0.1918	0.2405	-
Mushroom ^a	0.1062	0.2263	0.0151	-
Musty ^a	0.0030	0.0003	0.0000	-
Oak ^a	0.0592	0.1629	0.0004	-
Wet Wood/ Wet Sawdust	0.3874	0.4571	0.1598	-
Fresh Cut Wood	0.4286	0.2739	0.4403	-
Burnt Wood	0.0957	0.0043	0.0410	-
Rotten Wood	-	0.0028	0.0032	-
Yeast (Bread) ^b	0.0887	0.2089	0.4351	-
Vinegar ^a	-	-	0.0001	-
Bitter	-	-	-	0.0000
Sour/ Acid/ Tart	-	-	-	0.0000
Sweet	-	-	-	0.0000
Astringent	-	-	-	0.0000
Mouth Burn (localized perception)	-	-	-	0.0000
Warm (global feeling)	-	-	-	0.0740
Prickling	-	-	-	0.0001
Smooth	-	-	-	0.0000

Superscripts (a-c) indicate whether attribute is from ^aoutside, ^bmiddle, or ^cinner tier of Wine Aroma Wheel. ^dAttributes not presented in a sensory modality are indicated with a dash (-).

Table 3
 Final lexicon of the untrained panel (Panel U) showing p_{wine} from 2-way ANOVA

Attribute	Aroma Before Stirring	Aroma After Stirring	Flavor	Taste/ Mouthfeel
Apple ^a	0.0128	0.0036	0.0000	- ^d
Peach ^a	0.0293	0.4323	-	-
Melon ^a	0.0764	0.2743	0.0000	-
Pear	0.0210	0.0585	-	-
Lemon ^a	0.5981	0.3363	0.0127	-
Grapefruit ^a	0.0506	0.0052	0.0098	-
Pineapple ^a	0.6900	0.1904	0.0000	-
Rose ^a	0.0165	0.0413	0.1706	-
Green Bean ^a	0.0733	0.5398	-	-
Grape	-	-	0.0000	-
Asparagus ^a	0.6106	0.1990	0.0211	-
Cloves ^a	-	-	0.0000	-
Cut Grass ^a	0.1448	0.7807	-	-
Mushroom ^a	0.3281	0.3056	0.0076	-
Earthy ^b	0.0417	0.3941	0.0000	-
Alcohol ^a	0.4657	0.4144	0.0000	-
Pungent (Irritation from Alcohol) ^c	0.0418	0.1966	0.0000	-
Nutty ^b	0.2322	0.1428	0.0004	-
Honey ^a	0.1169	0.0022	-	-
Caramel ^a	0.0044	0.0721	0.0000	-
Raisin ^a	-	-	0.0000	-
Smoky ^a	-	-	0.0000	-
Vanilla ^a	0.8331	0.4013	-	-
Resinous/ Terpenes ^a	0.0574	0.0583	-	-
Oak ^a	0.1650	0.1567	0.0000	-
Cedar ^a	0.0027	0.0001	-	-
Medicinal	0.2422	0.5369	0.0000	-
Black Pepper ^a	0.8079	0.1214	0.0000	-
Vinegar ^a	-	-	0.0000	-
Sweet	-	-	-	0.0000
Sour/ Acid/ Tart	-	-	-	0.0000
Bitter	-	-	-	0.0000
Astringent	-	-	-	0.0000
Burn/Hot	-	-	-	0.0000
Cooling	-	-	-	0.0000
Smooth	-	-	-	0.0000

Superscripts (a-c) indicate whether attribute is from ^aoutside, ^bmiddle, or ^cinner tier of Wine Aroma Wheel. ^dAttributes not presented in a sensory modality are indicated with a dash (-).

Table 4
 Normalized RV coefficients for Panel T and Panel U for aroma before stirring, aroma after stirring, flavor, and taste/mouthfeel

		Panel T				Panel U			
		ABS ^a	AAS ^b	FLA ^c	TMF ^d	ABS	AAS	FLA	TMF
Panel T	ABS
	AAS	9.5
	FLA	3.8	4.6
	TMF	2.1	2.1	10.9
Panel U	ABS	1.8	3.5	1.5	0.2
	AAS	0.4	1.3	1.4	0.5	1.8	.	.	.
	FLA	2.7	2.8	11.4	11.6	1.4	1.3	.	.
	TMF	2.4	2.4	10.9	11.8	0.5	1.0	12.8	.

^a Aroma before stirring the wine glass.

^b Aroma after stirring the wine glass.

^c Flavor.

^d Taste/mouthfeel.

Table 5
 Multivariate Panel Performances from MANOVA on ANOVA-selected attributes

	Panel T		Panel U	
	F-approx. of Hotteling-Lawley trace	Number of significant canonical variates	F-approx. of Hotteling-Lawley trace	Number of significant canonical variates
aroma before stirring	1.93	3	1.68	4
aroma after stirring	1.78	4	2.06	3
flavor	1.76	3	2.00	2
taste/mouthfeel	3.25	1	4.57	3

Table 6
 Attribute-based performance measures of wine Panels T, U and D

Total attributes	Panel T	Panel U	Panel D
	110	76	130
Discriminated at $p < 0.05$	60	40	31
% discriminated	55	53	24
Disagreement attributes	5	9	43
% disagreement	4	12	33
Attributes in common	22	22	22
Discriminated at $p < 0.05$	15	12	11
% discriminated	68	55	50
Disagreement attributes	0	2	4
% disagreement	0	9	18

Table 7
 Frequency counts of number of feedbacks provided to panels and number of times the evaluation mean fell within the training target for Panel T and Panel U

Training Session	Panel T			Panel U		
	Number of feedbacks	Average training target size	Percentage evaluation mean fell within training target	Number of feedbacks	Average training target size	Percentage evaluation mean fell within training target
6	42	33.4	64.3	99	19.6	63.6
7	105	47.7	71.4	114	18.0	70.2
8	262	32.9	49.2	116	13.8	66.4
9	184	14.8	63.6	100	13.0	75.0
Total	593	29.9	58.7	429	16.1	68.8

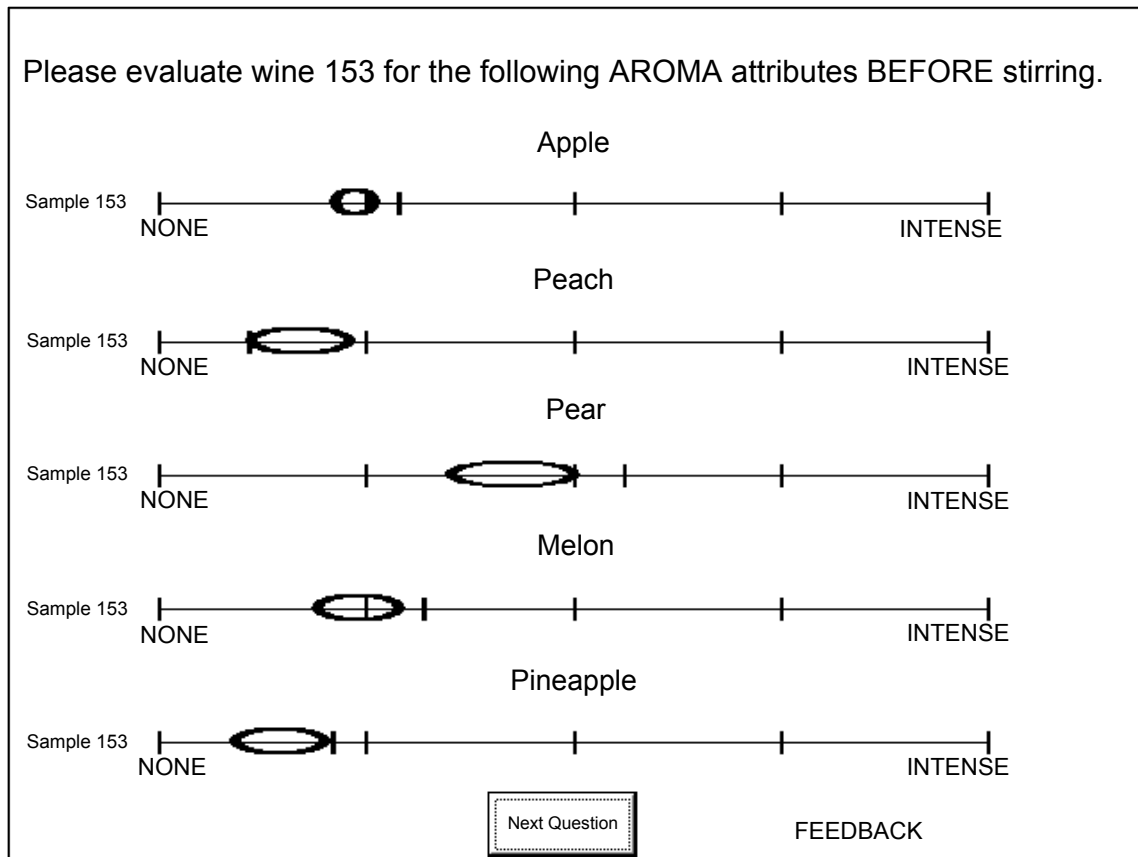
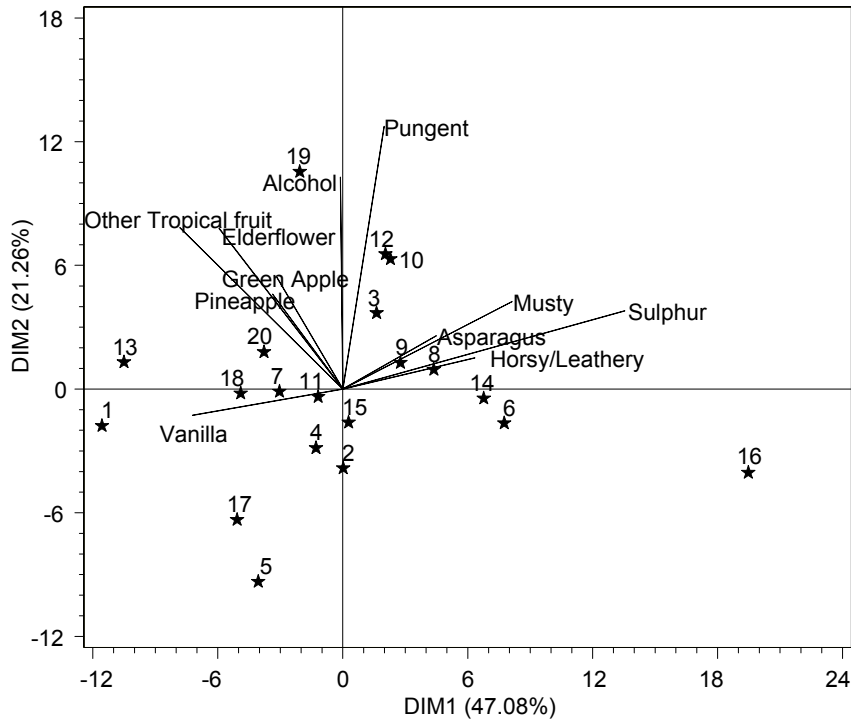


Fig. 1. During training sessions, immediate feedback was displayed to panelists following selected line scale questions. This black and white representation of the screen on which feedback was provided using ellipses, which in this study were based on the panel's own 90% confidence intervals around the mean value calculated from a previous training session. The thicker line is the panelist mark, which remains on the screen to allow for comparison with the ellipses. The reinforcement of attribute training targets is a central part the Feedback Calibration Method.

(a) Panel T



(b) Panel U

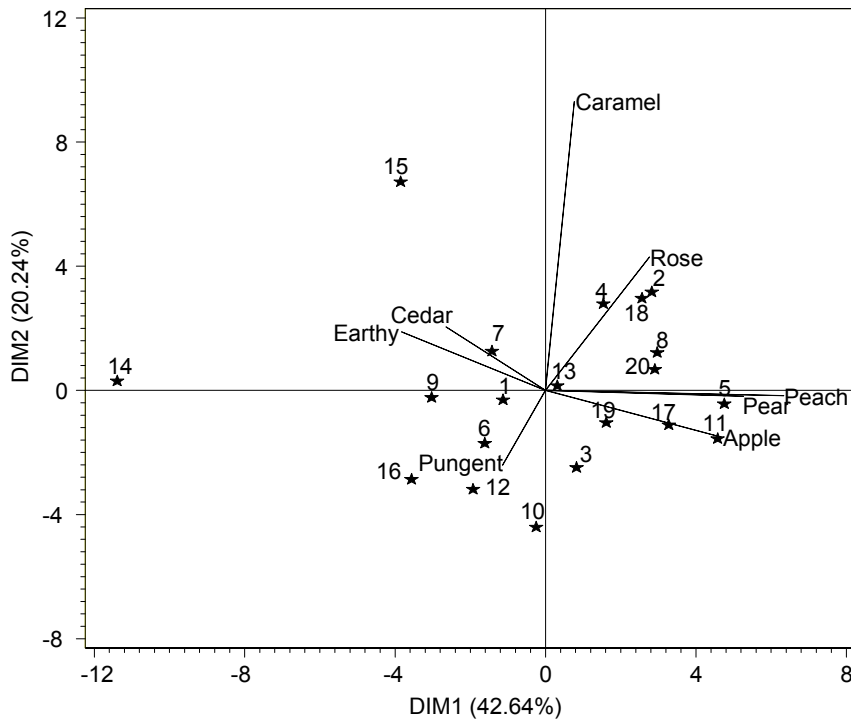
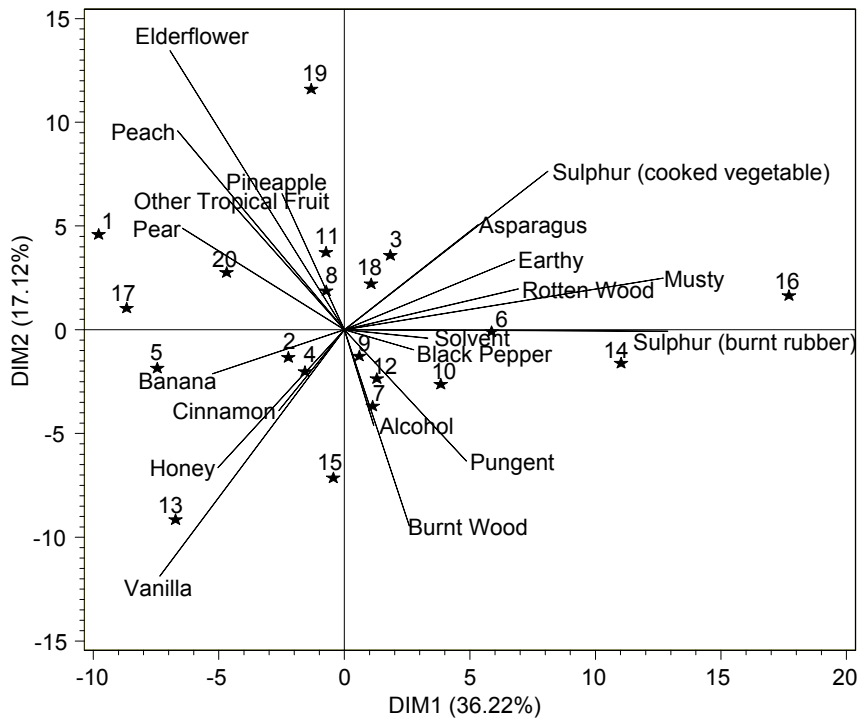


Fig. 2. Dimensions 1 and 2 of cov-PCA biplots run on the panel means for (a) Panel T and (b) Panel U for 20 wines and panel-selected aroma before stirring attributes. Stars indicate products and vectors indicate attributes. Similarity between these two configurations was slight (NRV=1.8).

(a) Panel T



(b) Panel U

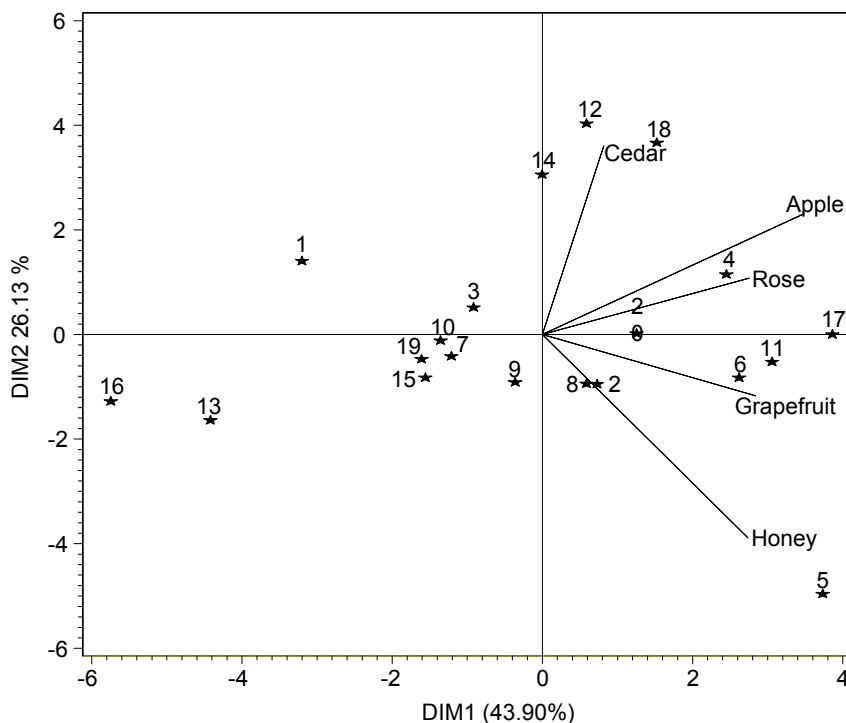


Fig. 3. Dimensions 1 and 2 of cov-PCA biplots run on the panel means for (a) Panel T and (b) Panel U for 20 wines and panel-selected aroma after stirring attributes. Stars indicate products, and vectors indicate attributes. Similarity between these two configurations was negligible (NRV=1.3).

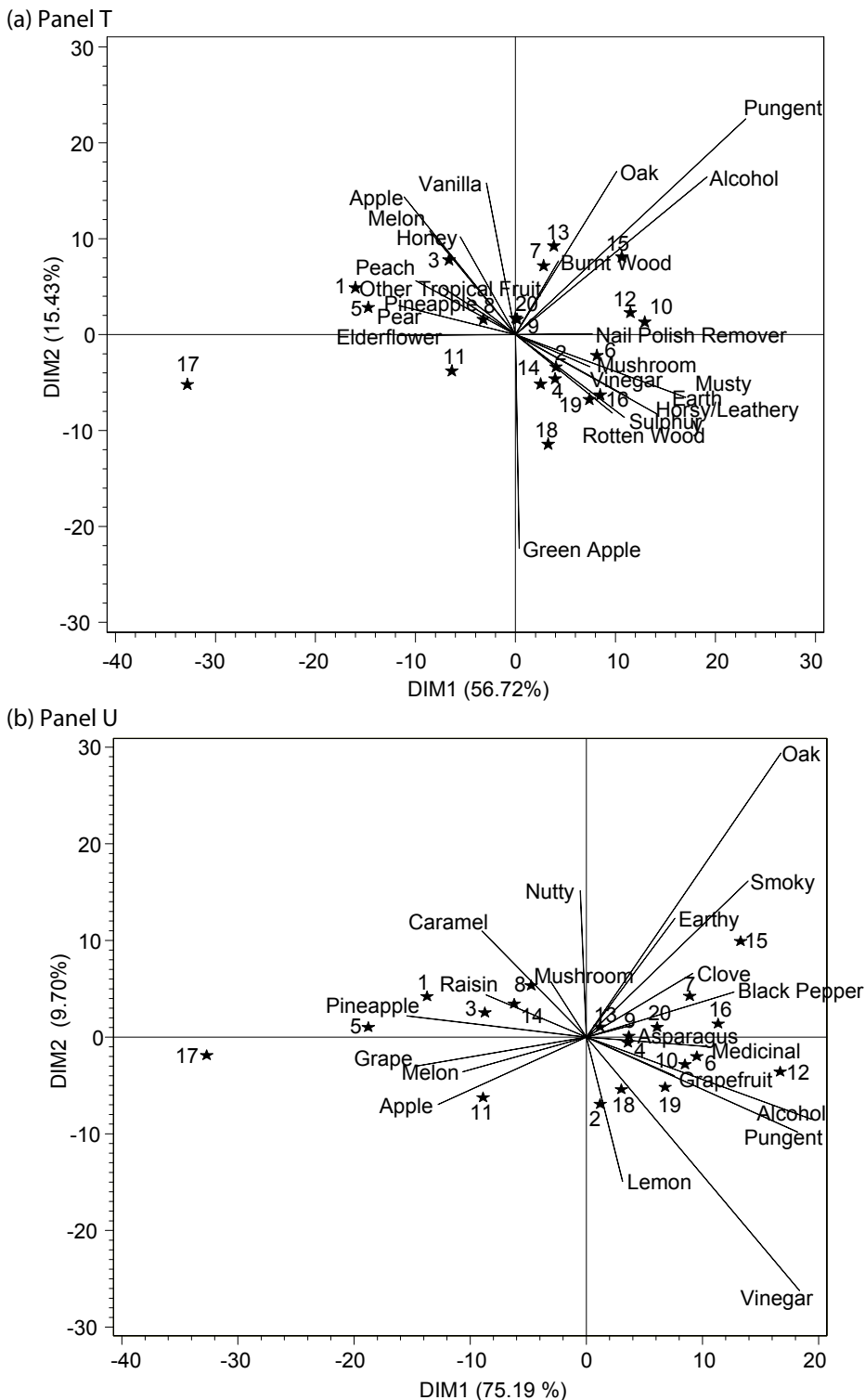
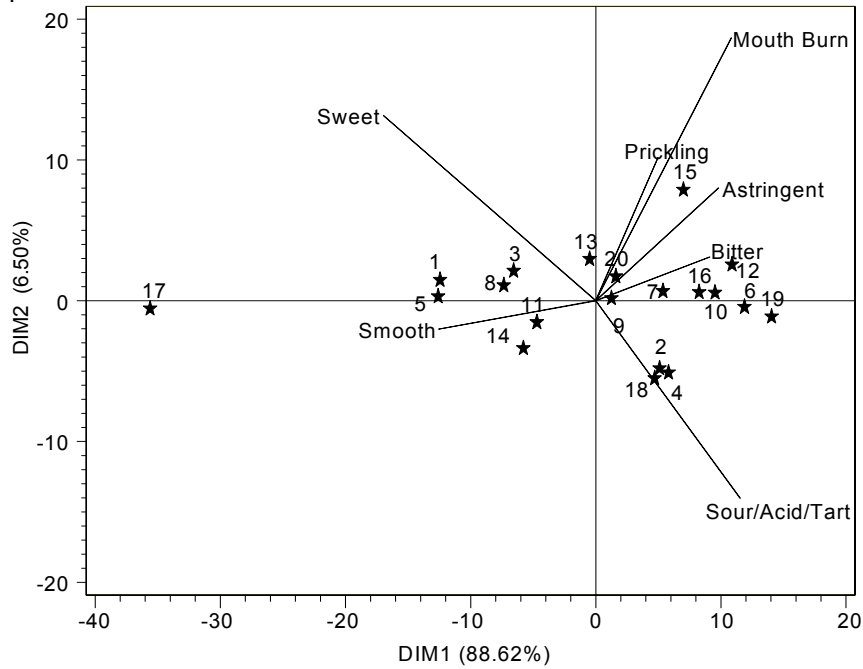


Fig. 4. Dimensions 1 and 2 of cov-PCA biplots run on the panel means for (a) Panel T and (b) Panel U for 20 wines and panel-selected flavor attributes. Stars indicate products, and vectors indicate attributes. Similarity between these two configurations was high (NRV=11.4). In part the higher NRV was due to the lower dimensionality, shown by the high first eigenvalue in the PCA.

(a) Panel T



(b) Panel U

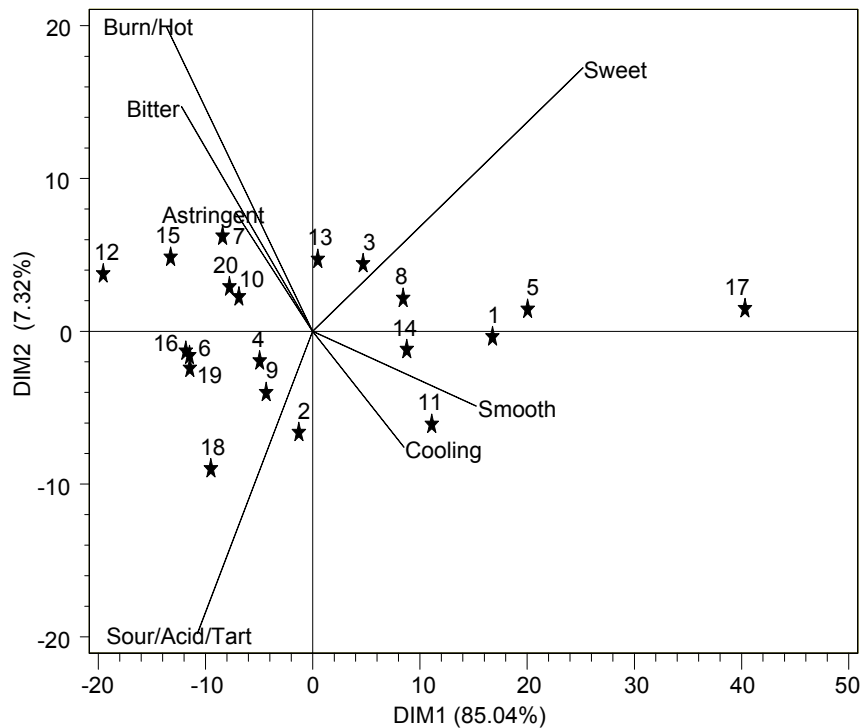


Fig. 5. Dimensions 1 and 2 of cov-PCA biplots run on the panel means for (a) Panel T and (b) Panel U for 20 wines and panel-selected taste/mouthfeel attributes. Stars indicate products, and vectors indicate attributes. These two configurations were highly similar (NRV=11.8). The higher NRV was partially due to the lower dimensionality, shown by the high first eigenvalue in the PCA.