Feedback calibration: A training method for descriptive panels

Chris J. Findlay[a,*], John C. Castura[a], Isabelle Lesschaeve[b,c]

[a] Compusense Inc., 679 Southgate Drive, Guelph, Ont., Canada N1G 4S2
[b] Cool Climate Oenology and Viticulture Institute, Brock University, St. Catharines, Ont., Canada L2S 3A1
[c] 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, 18*(2), 321-328.
Received 13 May 2005; received in revised form 18 February 2006; accepted 18 February 2006.
Available online 31 March 2006.

Abstract

Training targets were established using descriptive analysis profiles of 20 commercial red wines produced by a well-trained, experienced determination panel. After recruitment, screening and a basic sensory orientation of ten 2h common training sessions, 16 inexperienced panelists were divided by lottery into two panels. The control panel received a more conventional performance debriefing at the end of each training session. The experimental panel only received immediate graphical computerized feedback while in sensory booths. Both panels evaluated the same 20 wines and used the same scales and attributes. Panels were calibrated and responses compared to training targets. Performance was monitored daily as panels continued over a three-week period. Distance from target measurements showed similar improvement trends for both groups as measured by panelist and panel calibration. Results suggest the effectiveness of the feedback calibration method (FCM) in providing unbiased and effective training.

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

1. Introduction

Descriptive analysis is among the most powerful of the tools available to sensory scientists. Regardless of the approach being used to analyze the sensory attributes of products, descriptive panels require extensive training before the panel can become a reliable sensory instrument. In standard practice a conventionally trained panel receives feedback from the panel leader in a group setting after each training session. This often takes the form of a summary report that lists means and standard deviations for both the individual panelists and the panel as a whole. This may be provided to the panelists at the conclusion of a testing session (Meilgaard, Civille, & Carr, 1999). Kuesten, McLellan, and Altman (1994a, 1994b) found that computerized graphical feedback, provided to panelists in the form of a graph at the end of each training session, was effective in reducing context effects.

Improvements in the effectiveness of training, whether through time savings or decreases in resources invested in the panel, has practical implications for organizations that conduct or intend to conduct descriptive analysis. The role that timing of feedback has on performance has been investigated outside the field of sensory science, particularly in educational exercises (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Kluger & DeNisi, 1996; Kulhavy & Wager, 1993). This study examined the impact of a novel method of immediate computerized feedback, referred to as the feedback calibration method (FCM) on the training and calibration of a descriptive panel.

doi:10.1016/j.foodqual.2006.02.007

2. Materials and methods

2.1. Selection and storage of products

Red wine is a complex real-world product, making it ideally suited for use in testing the experimental panel training methodology. Vincor International Inc. (Mississauga, Ont., Canada) donated 20 commercial products in two lots. Two groups of wine were used for training and verification testing respectively. Group A wines (coded 1, 2, 3, 10, 13, 14, 16, 17, 18, 20) consisted of the following varietals: Merlot (3), Cabernet Sauvignon, Cabernet-Merlot, Cabernet Franc, Cabernet, Gamay, Pinot Noir, and Travigne. Group B wines (coded 4, 5, 6, 7, 8, 9, 11, 12, 15, 19) consisted of the following varietals: Merlot (2), Cabernet Sauvignon (4), Shiraz, Cabernet-Shiraz, and blended varietals (2). Wines were stored upside down in their cases in a climate controlled dark room maintained at 20 ±1 °C.

2.2. Sample presentation

Wine bottles were opened and decanted into 1.5L glass carafes 1h prior to training sessions. After wines were verified free of cork taint by the panel leader, 30 mL samples were poured into clear INAO wineglasses labeled with 3-digit blinding codes and covered with Petri tops and held at about 20 °C for 15 min prior to the training session. Samples were presented monadically to panelists via passthroughs. Phillips 100 W/115-125V PAR colortone red lights were used to mask wine color in the sensory booths. Between samples, panelists were instructed to clear their palates using a 5g/L pectin solution (Colonna & Noble, 2002) and rinse their mouths with distilled water at room temperature. Unsalted soda crackers were also available to panelists as palate cleansers, as was hot water to allow panelists to humidify their nasal passages.

2.3. Establishment of training targets

Eleven panelists from a standing external descriptive panel were invited to participate in a red wine determination panel (Panel D) and were paid for their involvement. During ten 2h training sessions, descriptors were elicited from the assessors using natural products, physical standards, and product samples, and a lexicon was developed. Scale usage exercises were conducted to establish intensity targets on a line-scale anchored at 0 and 100. All training sessions for the panel were recorded on videotape to provide a reference to assure consistency of methods used.

The panel first evaluated 10 wines (Group A) using a 10x10 Williams' Latin square design with three replicates. Subsequently, they evaluated 10 different wines (Group B) using the same design with two replicates. They used 130 line-scale attributes in five categories: aroma before stirring, aroma after stirring, taste, mouthfeel, and flavor.

After analysis, training targets were established for a representative subset of 31 attributes. Selection was influenced by the need to avoid potential dumping effects that might arise if attributes that were key characteristics of the product category, or remarkably intense, were not part of the lexicon. Fourteen of the 31 attributes showed significant differences amongst the Group A wines at $\alpha = 0.1$, both according to results obtained from a two-way mixed-model analysis of variance (ANOVA) and from a repeated measures ANOVA (RANOVA). Panel D means for product · attribute represented discrete targets, from which Tukey's Honestly Significant Difference (HSD) at 5% significance for attributes for the Group A wines was added and subtracted to establish ranges for the training targets. Minimum or maximum values for the target ellipses were not permitted to extend beyond the endpoints of the scale.

2.4. Recruitment, selection and initial training

The following criteria were established for potential inexperienced panelists: non-smoker, available during testing times, drink red wine at least once per month, no previous experience in sensory

Please refer to the published manuscript in *Food Quality and Preference* for all quotes and citations
doi:10.1016/j.foodqual.2006.02.007

analysis, no health problems that might interfere with testing, and no specialized knowledge of wines. A two round screening procedure identified 16 panelists that were well-suited to performing descriptive tasks and eliminated respondents with sensory deficiencies, such as inability to identify basic tastes.

The selected panelists received common descriptive sensory training in ten 2h sessions over a two-week period. The wine attribute lexicon, based on the determination panel, was introduced (Tables 1–3). During the common training period, the inexperienced panelists discussed and proposed additional attributes. Prune aroma and tropical fruit aroma were incorporated into the ballot as line-scale questions, but no feedback was provided for these attributes since they were not part of the lexicon established by Panel D. To alleviate potential dumping effects, panelists had the opportunity to indicate the presence of 9 new aroma descriptors and 27 new flavor descriptors using a pick list. Other attributes could be indicated via comment questions. All common training sessions (1–10) of the inexperienced panelists were recorded on videotape to assure consistency of methods.

2.5. Panel assignment

Panelists were assigned by lottery to either the control panel (Panel C) or the experimental panel (Panel E), such that each panel was composed of five women and three men. Which group was to become Panel C and which was to become Panel E was determined by a coin toss. Panelists were told the group was split because 16 panelists could not be served in the sensory booths concurrently. Training sessions were staggered to further discourage interaction between panels (Table 4).

2.6. Experimental design

Panel E and Panel C commenced scale usage exercises in the sensory booths on Session 11. At each training session, five of the ten Group A wines were selected and presented according to a modified 5x5 Williams' Latin square design, such that each training wine was evaluated five times during the course of training (Table 5). Training consisted of 10 sessions over a three-week period. On the 15th, 18th, 21st, and 24th sessions, Panel E and Panel C evaluated 5 Group B wines according to a modified 5-by-5 Williams' Latin square design, such that each Group B was evaluated two times during the course of the study. Neither panel received any feedback when evaluating Group B wines (Fig. 1).

2.7. Training feedback methodologies

Feedback for Panels C and E was the same in content but differed in timing and setting. The number of times feedback was provided increased from session to allow panelists to become accustomed to receiving the information (Table 5).

Panel C received all feedback in more traditional debriefings at the conclusion of each training session. These debriefings were video recorded to maintain a comprehensive record of the issues raised. Feedback was presented to each panelist in the form of a numerical report that indicated the panelists' responses and the acceptable range. Reports were augmented with the panel leader's commentary and visual displays of the acceptable range on a line-scale shown using an overhead projector. Food and magazines were made available to Panel C after the debriefings.

All feedback to Panel E appeared immediately after each panelist evaluated all line-scale attributes on each screen, and took the form of ellipses that showed the acceptable ranges (Fig. 2). The panelist's responses remained visible for comparison when feedback was shown. Panelists had the option of re-evaluating a sample in the presence of feedback. The panel leader did not provide commentary to Panel E as there were no debriefing sessions. To balance potential motivation effects caused by group interaction, members of Panel E were instructed to remain in the training

doi:10.1016/j.foodqual.2006.02.007

room for at least 20 min at the end of the training session to socialize, but not to discuss the panel. They were permitted to remain for up to an hour. Food and magazines were made available to Panel E during this social time.

2.8. Data collection and analysis

Data obtained from Panel D were collected and analyzed using Compusense five (Release 4.4, 2002, Compusense Inc., Guelph, ON, Canada), which provide results of the two-way fixed-effects ANOVA of the ten Group A wines and Tukey's HSD ($\alpha = 0.05$) as well as the means of the ten Group B wines. Data obtained from Panel C were also collected using Compusense five 4.4. An enhanced version of Compusense five 4.4 (version 4.5.19) was used to collect data from Panel E because it was programmed to provide graphic feedback to panelists during the test.

All univariate and multivariate ANOVAs were conducted on data from Panel C and Panel E using the General Linear Model in SPSS (Release 9.0.1, 1999, SPSS Inc., Chicago, IL, USA). Tukey's HSD was used as a post-hoc test to separate means when differences were detected by univariate ANOVA at $\alpha = 0.1$.

Principal Component Analysis (PCA) was performed on the correlation matrix based on mean values for each attribute*wine using Senstools (Release 3.1.6, 2002, OP&P Product Research BV, Utrecht, The Netherlands). Senstools was used to conduct RANOVA. Classical generalized procrustes analysis (GPA) was performed in Senstools using translation around the mean and isotropic scaling.

To understand changes in the panel and their use of attributes during training, responses from Panels C and E from the 10-session training period were collapsed into a data matrix containing five presentations of each of the 10 Group A wines, together with Panel D's three replicates of the 10 Group A wines. The 10 wines were treated as sets, the 13 panel by replication as objects, and a single replicate.

Pearson's coefficient of correlation and distance from target calculations (Castura, Findlay,& Lesscheave, 2006) were performed in Excel 2002 (Release 10.6501.6626 SP3, Microsoft Corporation, Redmond, WA, USA).

3. Results and discussion

3.1. Panel improvement over time

At the time the panelists were divided by lottery into two panels, neither the individual's level of calibration nor their potential for improvement were known. Several approaches were used to assess panel improvement over time:
(i) GPA: GPA was run on the collapsed data set and produced an 8-dimensional multivariate space that accounted for 87.74% of real and 94.48% total variance. The first two dimensions both had eigenvalues greater than one and collectively accounted for 67.66% of real and 71.57% total variance. Panels C and E approach Panel D's position in multidimensional space (Fig. 3), indicating a growing consensus in attribute usage.
(ii) Pearson's coefficient of correlation: The observed responses were compared to the 165 discrete targets for each training session. Correlation between the actual and expected responses were evaluated and found to be similar for Panels C and E (Fig. 4).
(iii) Distance from target: Several distance-from-target measurements (Castura *et al.*, 2006) were used to assess the panel's responses relative to the training targets. The 8-member panels each provided 1240 responses per session, providing a very table per-observation distance from target. Panel C and Panel E improved at a similar rate, as demonstrated by the change in adjusted-distance-from-range during the training period (Fig.5).

Please refer to the published manuscript in *Food Quality and Preference* for all quotes and citations
doi:10.1016/j.foodqual.2006.02.007

3.2. Nature of feedback techniques studied

Conventional feedback is visual and auditory, as well as delayed and group-oriented. The panel leader can motivate and re-focus panelists with wandering minds or flagging motivation. Conventional feedback is idiosyncratic: the charisma, knowledge and communicative ability of the panel leader play a role in its effectiveness. By contrast, the experimental feedback is only visual. It is immediate and directed toward the individual assessor in the sensory booth. It is consistent and does not differentiate among panelists. The panelist's reaction to this form of feedback remains unknown because their immediate reaction to the feedback could not be monitored. However, at the end of the project, a review with Panel E was positive towards the use of the onscreen targets. Rather than being overwhelmed by the immediate feedback, the panelists commented that they would like to have feedback for all attributes and all samples.

3.3. Assessing the validity of the training targets

Training targets would be ineffective if they did not reflect the underlying sensory characteristics of the products. The consistent improvement of inexperienced panels during the study and their ability to discriminate wines are evidence of the validity of the training targets.

The attributes used by Panel D were substantially reduced when transferred to the inexperienced panelists. The large number of attributes evaluated by Panel D allowed a well-trained panel to describe the full complexity of the wine using a complete lexicon; it would have been difficult for the panel leader to know a priori which of the attributes would be used to discriminate wines. Problems that might arise due to attribute reduction were minimized by selecting attributes that discriminated wines or were characteristic of the product category.

4. Conclusions

The results of this preliminary study suggest that a panel trained in scale usage with immediate feedback reached a similar level of proficiency and improvement to a panel that was more conventionally trained. Immediate feedback may have the potential of supplementing the skills and experience of the panel leader, decreasing reliance on specialized and highly experienced panel leaders, and providing an efficient way for panel leaders of varying skill levels to calibrate and maintain descriptive sensory panels.

Further research into the combination of FCM with best practices in panel training will be conducted. The purpose will be to optimize the training process and assure proficiency of panelists and the panel.

Acknowledgements

The authors thank the National Research Council of Canada's Industrial Research Assistance Program (NRC-IRAP) for funding assistance provided to Project #474766, as well as Vincor International Inc. for providing the commercial wine used in this study. The authors also acknowledge the contributions of Amanda Bartel and Karen Phipps, who conducted panel recruitment, screening and training, and the sensory panelists, whose commitment made this study possible.

Please refer to the published manuscript in *Food Quality and Preference* for all quotes and citations
doi:10.1016/j.foodqual.2006.02.007

References

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., & Lesscheave, I. (2006). Monitoring calibration of descriptive sensory panels using distance from target measurements. *Food Quality and Preference, 17*, 282–289.

Colonna, A. & Noble, A. (2002). Protocol for evaluating astringent red wines in research and industry. In *Abstract published in proceedings of: Bacchus to the Future – The inaugural Brock university wine conference*. (p. 129).

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. P*sychological Bull, 199*(2), 254–284.

Kuesten, C. L., McLellan, M. R., & Altman, N. (1994a). Influence of computerized panel training on contextual effects. *Journal of Sensory Studies, 9*, 401–412.

Kuesten, C. L., McLellan, M. R., & Altman, N. (1994b). Computerized panel training: Effects of using graphic feedback on line scale usage. *Journal of Sensory Studies, 9*, 413–444.

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 Cliffs, NJ: Educational Technology Publications.

Meilgaard, M., Civille, G. V., & Carr, B. T. (1999). *Sensory evaluation techniques* (third ed.). Boca Raton, FL: CRC Press LLC.

doi:10.1016/j.foodqual.2006.02.007

Table 1
Aroma attributes transferred to inexperienced panels

Attribute	Definition – the aroma associated with
Floral	Floral white wine, elderflower, flowers
Rose	Rose petals, rose water
Earthy/Musty	Wet/dry soil, mushroom, wet clay, musty
Alcohol	Alcohol
Pungent	Burning sensation in the nasal cavity elicited by alcohol
Oak barrel	Oak, oak chips, oak barrel
Smoky	Smoke, mesquite, hickory smoke
Fermented	Newly fermented juice/wine
Currant	Red currant, black currant, cranberries
Cherry	Cherry pie filling, cherry, black cherry, red licorice
Sulphur	Canned corn, sauerkraut, hydrogen sulfide, rotten eggs
Black pepper	Fresh ground black pepper, pepper corns
Honey	Liquid honey, honey comb
Medicinal	Medicine
Grape	Concord grape juice, grapes
Red berries	Raspberries, strawberries

doi:10.1016/j.foodqual.2006.02.007

Table 2
Taste/mouthfeel attributes transferred to inexperienced panels

Attribute	Definition
Sweet	Taste on the tongue stimulated by sugars and high potency sweeteners
Sour	Basic taste on the tongue stimulated by acids
Bitter	Taste on the tongue stimulated by solutions of caffeine and quinine
Salty	Taste on the tongue stimulated by sodium salt, especially sodium chloride
Astringent	The chemical feeling factor on the tongue or other surfaces of the oral cavity described as puckering/dry, and associated with tannin or alums

Please refer to the published manuscript in *Food Quality and Preference* for all quotes and citations
doi:10.1016/j.foodqual.2006.02.007

Table 3
Flavor attributes transferred to inexperienced panels

Attribute	Definition – flavor associated with
Earthy/musty	Wet/dry soil, mushroom, wet clay, musty
Oak barrel	Oak, oak chips, oak barrel
Fermented	Newly fermented juice/wine
Alcohol	Alcohol
Pungent	In-mouth and retronasal burning sensation elicited by alcohol
Yeasty	Active dry yeast
Currant	Red currant, black currant, cranberries
Cherry	Cherry pie filling, black cherry, cherries, red licorice
Grape	Concord grape juice, grapes
Black pepper	Fresh ground black pepper, pepper corns

doi:10.1016/j.foodqual.2006.02.007

Table 4
Schedule for2h training sessions (T1–T10) for Panels C and E

	Panel C	Panel E
10:00	Scale usage calibration in sensory booths	
11:00	Delayed feedback in 1st floor training room (video-recorded group debrief session with panel leader)	Scale usage calibration and Immediate feedback in sensory booths
12:00		Social time in 3rd floor training room

Please refer to the published manuscript in *Food Quality and Preference* for all quotes and citations
doi:10.1016/j.foodqual.2006.02.007

Table 5
Schedule for the scale usage calibration of Panels C and E in the sensory booths

Session[a]	Feedbacks[b]	Wines Presented[c]				
T1	15	14	1	2	3	13
T2	20	13	14	20	17	16
T3	30	16	18	1	20	2
T4	30	1	10	17	2	14
EV1	0	5	8	12	15	19
T5	45	20	16	10	1	3
T6	60	17	3	18	14	20
EV2	0	19	5	12	8	15
T7	75	10	13	14	16	18
T8	90	2	20	3	13	10
EV3	0	4	6	7	9	11
T9	105	18	17	13	10	1
T10	135	3	2	16	18	17
EV4	0	11	4	7	6	9

[a] Training and evaluation sessions are labeled T1–T10 and EV1–EV4, respectively.
[b] Indicates the number of times feedback was provided to each panelist in the session.
[c] Indicates the 5 wines presented to each panelist according to a modified 5 x 5 Williams' Latin square design.

doi:10.1016/j.foodqual.2006.02.007

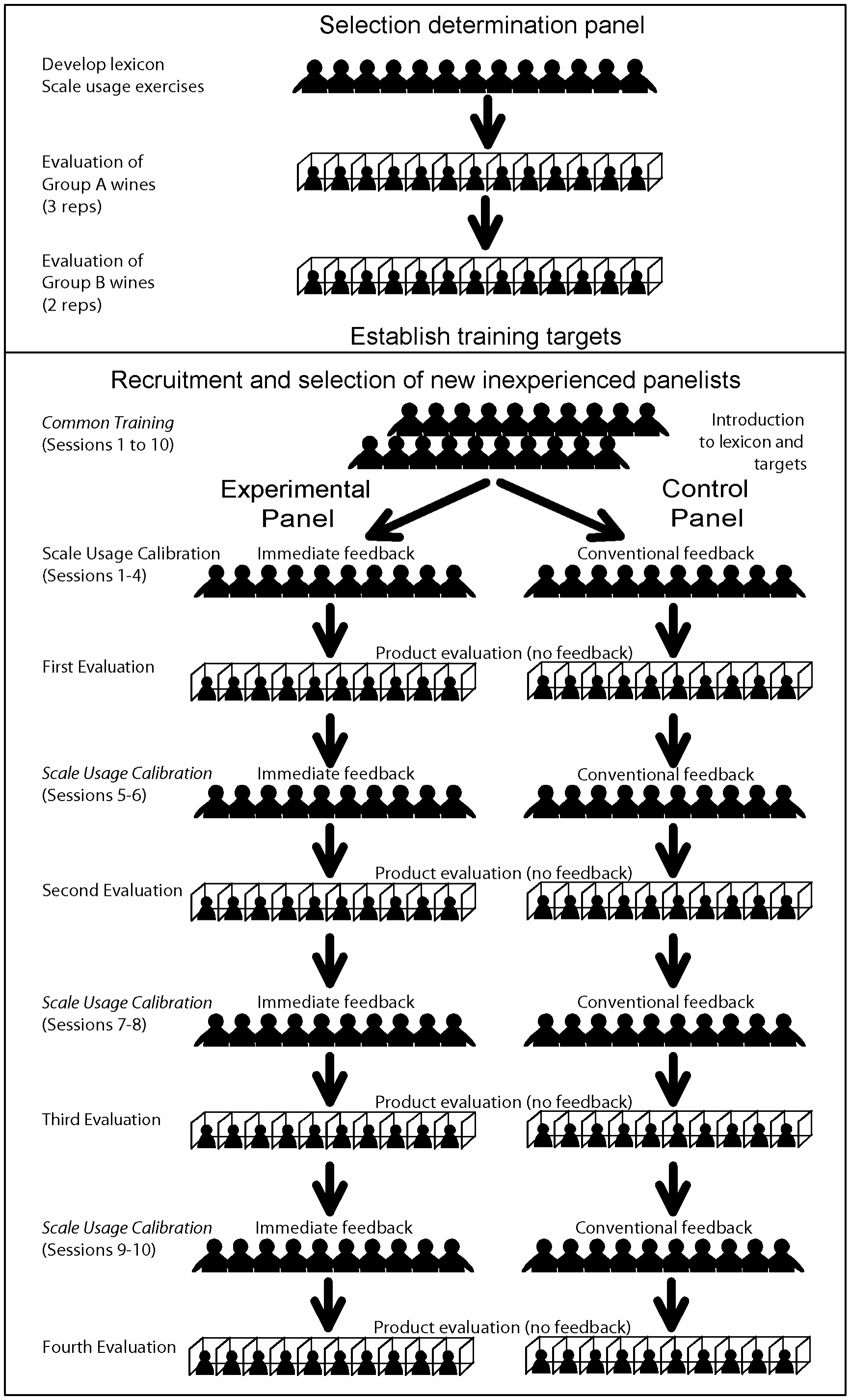

Fig. 1. Overview of the experimental plan.

Please refer to the published manuscript in *Food Quality and Preference* for all quotes and citations
doi:10.1016/j.foodqual.2006.02.007

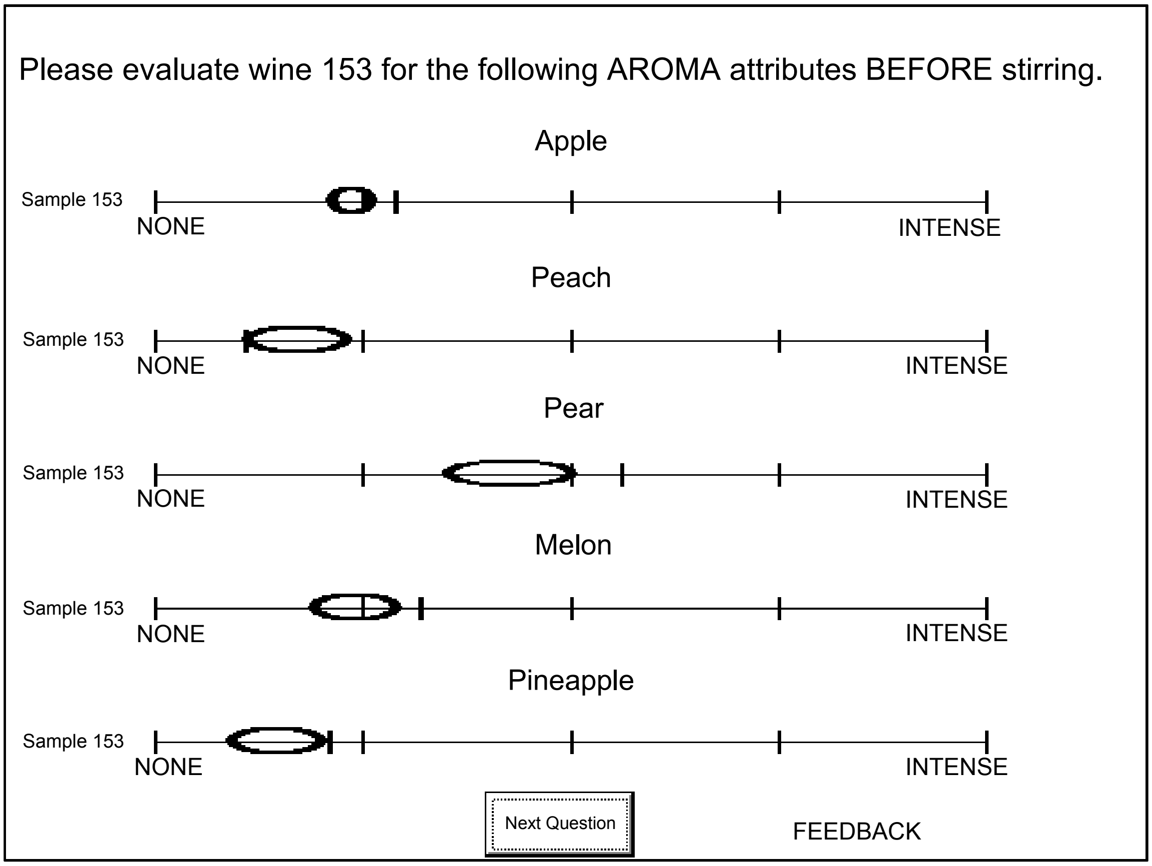

Fig. 2. An example of onscreen feedback in the form of target ellipses.

Please refer to the published manuscript in *Food Quality and Preference* for all quotes and citations
doi:10.1016/j.foodqual.2006.02.007

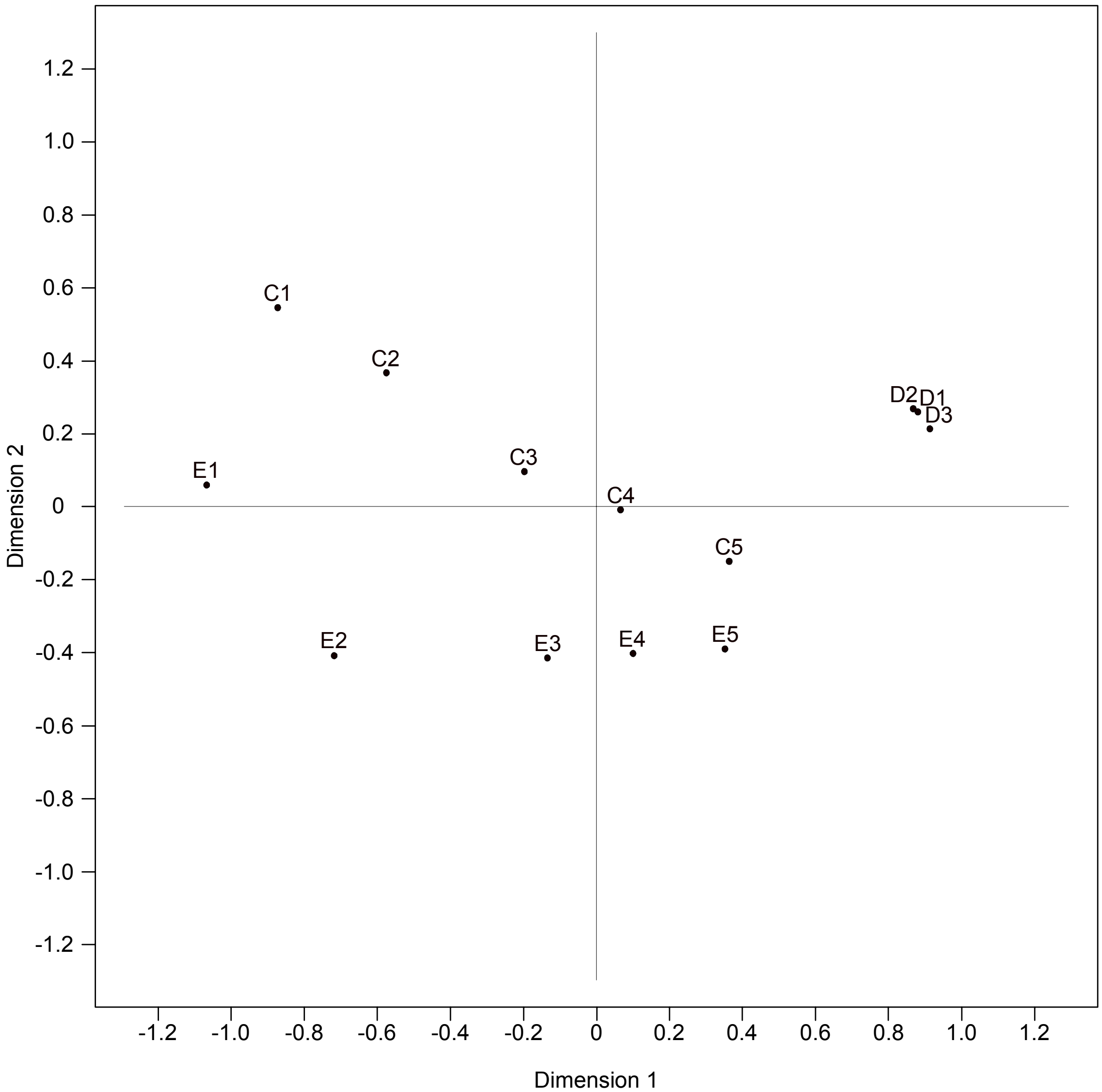

Fig. 3. Dimensions 1 and 2 of GPA group average showing results of Panels D, C, and E. Dimension 1 accounts for 54.63% of real variance. Dimension2 accounts for 13.03% of real variance. Data for Panel D were obtained from three replicates during evaluation (D1, D2, D3). Data for Panel C (C1–C5) and Panel E (E1–E5) were obtained from a collapsed data matrix from the 10 training sessions, and represent a change over time. Only the 31 attributes in the transferred lexicon were considered.

Please refer to the published manuscript in *Food Quality and Preference* for all quotes and citations
doi:10.1016/j.foodqual.2006.02.007

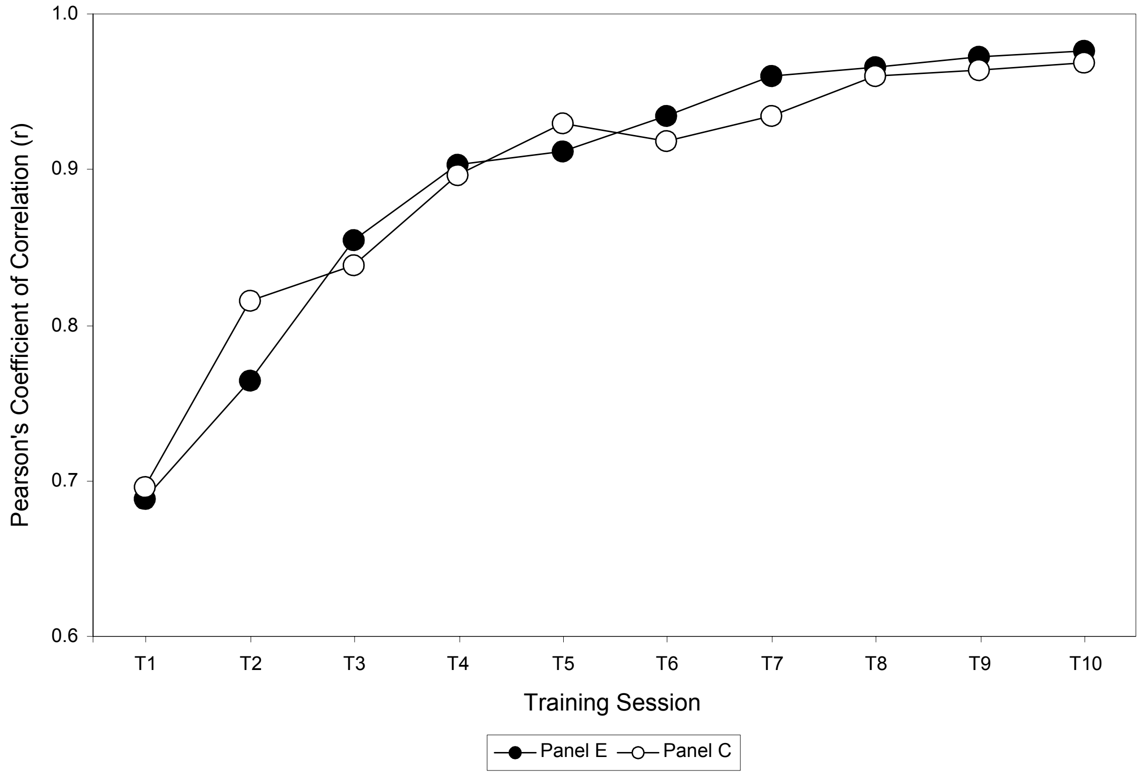

Fig. 4. Pearson’s coefficient of correlation between observed responses and expected discrete targets.

doi:10.1016/j.foodqual.2006.02.007

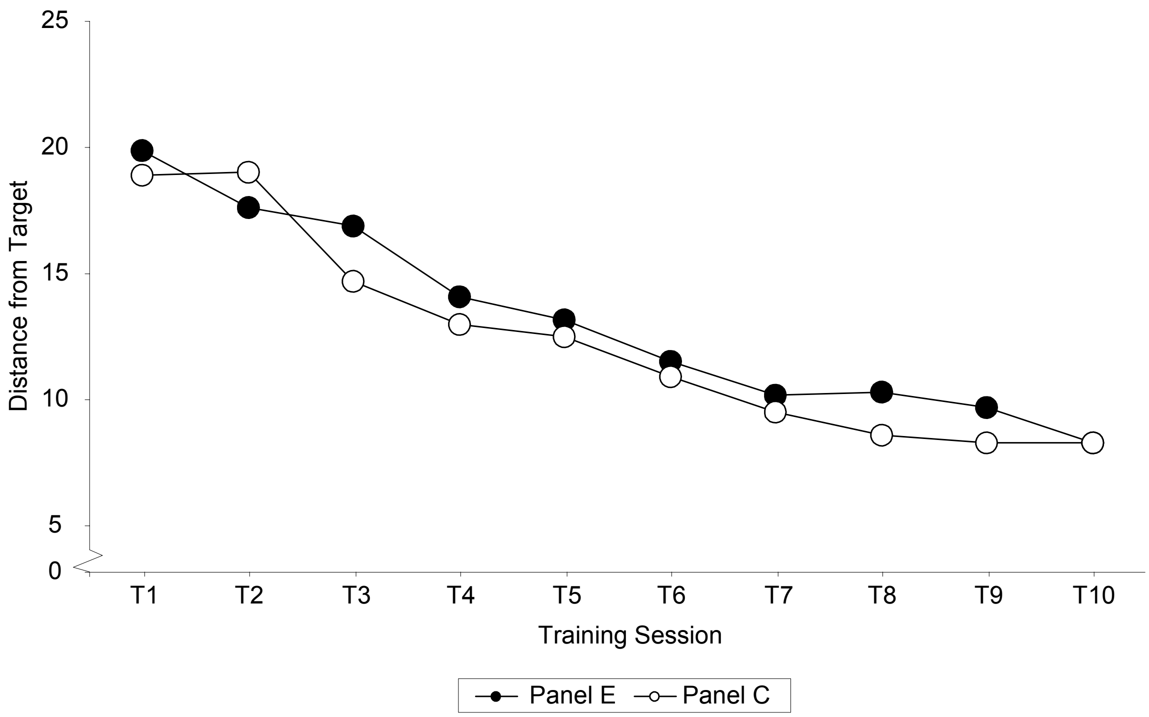

Fig. 5. Adjusted distance from range (ADR) for Panels C and E over the 10 training sessions.

doi:10.1016/j.foodqual.2006.02.007