Monitoring calibration of descriptive sensory panels using distance from target measurements

John C. Castura[a,*], Chris J. Findlay, 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: jcastura@compusense.com (J.C. Castura)

Originally published in *Food Quality and Preference, 16*(8), 682-690.
Reprinted in *Food Quality and Preference, 17*(3-4), 266-276.
Received 15 September 2004; accepted 20 March 2005. Available online 13 May 2005.

Abstract

Training targets can be established from product profiles that provide an objective representation of the underlying sensory characteristics of a group of products. If available to a panel leader, these training targets can be used to calibrate a panel and measure the accuracy of their responses. Response accuracy can be determined either by frequency counts—how often the target was hit versus how many opportunities there were to hit it—or by distance from target measurements—which attempt to further quantify the degree to which the target was hit or missed. In addition to the frequency counts, four distance-from-target methods are presented and discussed—Distance from Target, Distance from Range, Adjusted Distance from Target, and Adjusted Distance from Range—each of which provides insights into the degree to which the panel is calibrated.

Keywords: Calibration; Descriptive; Training; Target

1. Introduction

The people that comprise the descriptive sensory panel are the detectors of the analytical instrument of descriptive analysis. When products are evaluated by a well-trained descriptive panel using established sensory methodologies, the panel leader can expect to obtain reliable information about products undergoing study. Screening, selection, training, and panel maintenance are exercises that help the panel attain proficiency prior to product evaluation. Some of the statistical analyses that provide information about panel proficiency include analysis of variance, multiple comparison tests, principal component analysis, and generalised procrustes analysis (Lawless & Heymann, 1998), and are often calculated in the absence of a ''true value'' that is known to the sensory scientist. When developing procedures to assess panel proficiency, the ProfiSens project participants avoided the thorny issue of what constitutes a ''true value'' by employing the concept of an ''expected result'' (McEwan, 2000). The approaches to measuring panel calibration outlined in this paper were developed in response to data collected from descriptive panels that were calibrated to established and fixed training targets in a recent study. It was against these training targets, which constituted the expected values, that panel calibration and panellist calibration were then measured.

The authors conducted an experiment in which a lexicon and intensity targets were established for 20 red wines by a well-trained descriptive panel (Findlay, Castura, & Lesschaeve, 2003a). Subsequently two new panels, each composed of 8 inexperienced panellists, were trained to identify selected attributes, and calibrated using the training targets established by the determination panel (Panel D). The panels evaluated 20 products using 31 attributes for which

Please refer to the published manuscript in *Food Quality and Preference* for all quotes and citations (doi:10.1016/j.foodqual.2005.03.011).

intermittent feedback was provided. When a panellist uses a line scale to evaluate a product for an attribute, the observed response can be compared to the target, which is the expected value of the response (Findlay *et al.*, 2003a; Findlay, Castura, & Lesschaeve, 2003b). The improvement of the inexperienced panels over the five-week training period was monitored using in-range frequency counts and four approaches to calculating distance from target.

In-range frequency counts provide a simple yet meaningful indication of the level of training of both a panel and the individual panellists. The number ''hits'' and ''misses'' can be assessed on a per-product, per-attribute, and overall basis, providing the panel leader with useful information about the analytical instrument of descriptive analysis and the detectors that comprise it. The frequency counts provide an indication of whether or not the panel or panellists are reaching the training targets; distance from target approaches attempt to quantify the degree to which responses are hitting the target. Measurements discussed are Distance from Target, Distance from Range, Adjusted Distance from Target, and Adjusted Distance from Range.

2. Selecting a training target

When calibrating a descriptive panel using training targets, panellists evaluating a product for a line scale attribute can be provided with feedback in the form of a discrete point on the line scale. Alternatively, feedback to panellists can be in the form of an acceptable range. Variation of product perception provides justification for the latter approach.

If K panellists on a well-trained determination panel evaluated N products for M attributes, Tukey's Honestly Significant Difference (HSD) can be calculated for each of M attributes. Tukey's HSD contains variance introduced by the judges, variance among samples, and other factors. Tukey's HSD is a conservative multiple sample comparison procedure, and its relatively large range reduces the risk of declaring spurious differences between products. If the panel had undergone similar training on all attributes, the size of Tukey's HSD for the *j*-th attribute will be related to the difficulty of that attribute in the context of the N products being studied. Training targets established using the N product profiles of a well-trained panel can be envisioned as N acceptable ranges on a line scale. When acceptable ranges are well separated with relatively little overlap and a panel is trained to reach these training targets, their responses could be expected to differentiate the products. When the acceptable ranges are overlapped with large range sizes, differentiating the products will prove to be more challenging. A different product category, or even different products within the product category, may create a context in which the *j*-th attribute becomes more or less difficult to scale.

It should be noted that just because products are formulated differently does not guarantee that panellists will detect a sensory differences; the sensory attribute in the product context may fall within a range that cannot be differentiated. This idea forms the basis for psychophysical investigations into the Just Noticeable Differences (Lawless & Heymann, 1998). The ability of a panel to differentiate N products for the *j*-th attribute may say as much or more about the *j*-th attribute as the calibration of the panel.

When calibrating a descriptive panel, the discrete training target could be the average for K panellists for the *i*-th product and the *j*-th attribute. To create a training range, a discrete point can be bounded on either side by Tukey's HSD for the attribute, with minimum and maximum points occurring not beyond the line scale endpoints. The *k*-th panellist evaluating N products for the *j*-th attribute is also comparing multiple samples. When an observed response O_{ijk} has missed the expected discrete response E_{ijk}, the response may still be considered "valid" if the response falls within the training range. In this case, the panellist's response is judged to be indistinguishable from that of a calibrated panellist.

(doi:10.1016/j.foodqual.2005.03.011).

A possible source from which to obtain training targets is previously collected data from a well-trained determination panel. The same training targets used to provide feedback to panellists can also be used as a benchmark against which panel calibration can be measured. The quality of distance from target measurements depends entirely on the degree to which the training targets reflect the underlying sensory characteristics of the products.

3. Frequency counts and percentages of observations in range

Perhaps the most straightforward approach to determine whether the responses of a panellist or panel are within the acceptable range is to count the number of responses that fall within the training target (a ''hit'') versus the number of opportunities for responses to be within the training targets. This proportion of hits to opportunities can be expressed as a percentage to allow comparisons. There are several ways that the frequency counts and corresponding percentages can be calculated.

By attribute

3.1.1. By attribute, by product

For each attribute, products are listed, along with the training target for the product*attribute. The frequency count of in-range responses across all panellists, as well as the number of opportunities to hit training targets, can be listed. A listing of panellist codes and the raw data response can follow, to provide perspective on the degree to which individual panellists were out of range.

3.1.2. By attribute, by panellist

For each attribute, panellists are listed along with the number of observations for which panellists were in range and the number of opportunities the panellists had to hit training targets. This information can also be expressed as a percentage.

3.2. Amount of feedback provided

Panellists can be overwhelmed if presented with too much feedback. Training targets can be available but not always presented to panellists. Tracking the quantity of feedback presented to panellists can be done by obtaining frequency counts of the number of feedbacks both per-attribute and overall. Frequency counts of the number of training targets can also be obtained. This provides the panel leader with a perspective on whether it is possible to increase the number of feedbacks, if so desired.

3.3. Overall in range by attribute

The number of observations in range across all panellists and products can be determined on an attribute-by-attribute basis. By looking at the number of opportunities available to hit training targets for each attribute, a percentage can be calculated that indicates how well the panel was able to ''hit'' the training targets. An overall frequency of hits, opportunities, and percentage of hits can be calculated to indicate the ability of the panel to hit the training targets overall. This information can be used by the panel leader to determine whether the panel requires additional training and on which attributes additional training is required.

3.4. Overall in range by panellist

The number of observations in range across all attributes and products can be determined on a panellist-by-panellist basis. By looking at the number of opportunities available to each panellist to hit training targets, a percentage can be calculated that indicates how well each panellist was able

Please refer to the published manuscript in *Food Quality and Preference* for all quotes and citations
(doi:10.1016/j.foodqual.2005.03.011).

to "hit" the training targets. This information can be used by the panel leader to determine which panellists require additional training. Cross-referencing this information with the by-attribute, by panellist data can indicate the attributes with which the panellist is having the most difficulty.

4. Four approaches to distance from target

4.1. Distance from target (DT)

This approach compares panellist data to the discrete target within the range (Fig. 1), which can be calculated in several ways depending on the purpose for the comparison:

- across all panellists, products and attributes, if the focus is on overall panel calibration,
- by panellist to determine whether one or more panellists requires additional training,
- by attribute to determine whether particular attributes require additional training.

Other combinations are possible.

DT is calculated by taking the absolute value of the differences between the observed and the expected response, so it will also be denoted as |DT|. There are several approaches possible, some of which are outlined below.

The panel leader can monitor the calibration of a panel by calculating |DT| for N products, M attributes, and K panellists.

$$|DT|_{panel} = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{K} \left(|O_{ijk} - E_{ij}| \right) \quad (1)$$

Two panels, each comprised of the same number of panellists, evaluating the same products using the same attributes, can be compared using |DT|. This measures the calibration of each panel panel directly and monitors their relative improvement. Accounting for missing responses in the data set can be handled by finding |DT|..., the average per-observation DT.

$$|DT|_{...} = |DT|_{panel} / n \quad (2)$$

where n is the number of observations. In a data set with no missing values, $n = K \cdot M \cdot N$.

Alternatively, if there is no missing data but the panels are comprised of differing numbers of panellists, the |DT| of the smaller panel can be grossed up to account for the differences. If Panel A consists of 10 panellists and Panel B consists of 13 panellists, it is possible to compare the panels by comparing 1.3 * $|DT|_A$ to $|DT|_B$.

The panel leader can determine the performance of the panel for the *j*-th attribute by calculating |DT| for N products and K panellists.

$$|DT|_j = \sum_{i=1}^{N} \sum_{k=1}^{K} \left(|O_{ijk} - E_{ij}| \right) \quad (3)$$

The panel leader can determine the performance of the panel

Please refer to the published manuscript in *Food Quality and Preference* for all quotes and citations (doi:10.1016/j.foodqual.2005.03.011).

for the *k*-th panellist by calculating
|DT| for N products and M attributes.

$$|DT|_k = \sum_{i=1}^{N} \sum_{j=1}^{M} \left(|O_{ijk} - E_{ij}| \right) \tag{4}$$

4.2. Distance from range (DR)

This approach compares the panellist response to the acceptable range (Fig. 2). DR is a summation of the difference between responses that are out of the acceptable range and the closest responses that fall within range. When the *k*-th panellist's response falls within the acceptable range for the ith product for *j*-th attribute, the response is held to be consistent with a well-calibrated response, adding zero to DR. The range (r) can be established on a per-attribute basis (r_j), for example using Tukey's HSD, or on a product *attribute basis (r_{ij}), for example using 90% confidence intervals. Theoretically each panellist could have a range that reflects their own scale usage patterns (r_{ijk}), assuming that statistical techniques would later be used to compensate for differences in scale usage among panellists; however, in practice when large numbers of products and attributes are used, this approach could be quite onerous for the panel leader to manage.

Absolute distance from range (|DR|) for K panellists who have evaluated N products using M attributes can be calculated in the following manner:

$$|DR|_{panel} = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{K} \left(|DR_{ijk}| \right) \tag{5}$$

where $DR_{ijk} = (|O_{ijk} - E_{ij}| - r_{ij})$ if $(|O_{ijk} - E_{ij}|) > r_{ij}$, and $DR_{ijk} = 0$ if $(|O_{ijk} - E_{ij}|) <= r_{ij}$. (6)

Accounting for missing responses in the data set can be handled by finding |DR|..., the average per-observation DR.

$|DR|... = |DR|_{panel} /n$ (7)

where n is the number of observations. In a data set with no missing values, $n = K \cdot M \cdot N$.

4.3. Adjusted Distance from Target (ADT)

This approach expresses DT as a ratio of the training target, thus creating a unitless measurement. Furthermore, ADT can be used when the panel leader recognizes that there are different levels of tolerance for missing targets. When ranges are wide, it could imply that the attribute is difficult; when ranges are narrow, it could imply that panellists can be well calibrated for that attribute. ADT divides DT by the acceptable range r prior to each summation (Fig. 3). When a panellist's response falls outside the expected range, the penalty is greater when the range is smaller.

The panel leader can monitor the calibration of a panel over time by calculating ADT for N products, M attributes, and K panellists for each session.

Please refer to the published manuscript in *Food Quality and Preference* for all quotes and citations (doi:10.1016/j.foodqual.2005.03.011).

$$|ADT|_{panel} = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{K} \left(|O_{ijk} - E_{ij}| / r_{ij} \right) \quad (8)$$

This calculation can be used to determine the average ADT for a single observation.

$$|ADT|_{...} = |ADT|_{panel} / n \quad (9)$$

where n is the number of observations. In a data set with no missing values, $n = K \cdot M \cdot N$.

$|ADT|_{...}$ can be used to determine level of calibration for panels comprised of different numbers of panellists that evaluate the same N products using the same M attributes. It may be possible to use $|ADT|_{...}$ to make comparisons among attributes for the same product, or among descriptive panels evaluating products in different product categories, but the results of several empirical studies will be required to determine how effectively $|ADT|_{...}$ compensates for the difficulty level of attributes.

4.4. Adjusted Distance from Range (ADR)

This approach treats each expected value as a range rather than as a discrete point (Fig. 4). ADR expresses DR as a ratio of the training target, thus creating a unitless measurement. Furthermore, ADR compensates for the expected difficulty level of each attribute. $|DR|_{panel}$ can be calculated on a session-by-session basis during panel training to monitor the change in panel calibration.

$$|ADR|_{panel} = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{K} \left(|DR_{ijk}| / r_{ij} \right) \quad (10)$$

where $DR_{ijk} = (|O_{ijk} - E_{ij}| - r_{ij})$ if $(|O_{ijk} - E_{ij}|) > r_{ij}$, and $DR_{ijk} = 0$ if $(|O_{ijk} - E_{ij}|) <= r_{ij}$. (11)

ADR can also be reduced to a per-observation average.

$$|ADR|_{...} = |DR|_{panel} / n \quad (12)$$

where n is the number of observations. In a data set with no missing values, $n = K \cdot M \cdot N$.

Both ADT and ADR approaches assume that attributes vary in their difficulty levels within a subset of products within a product category, and that this difference is well reflected in statistical measurements for conducting multiple comparisons. Tukey's HSD will be relatively low for attributes that are relatively easy to scale and relatively high for attributes that are relatively difficult to scale in their respective contexts.

The Beidler model plots a relationship between response, maximal response, and stimulant concentration. This plot takes on a sigmoidal shape that follows the underlying neurophysiological response (Meilgaard, Civille, & Carr, 1999). The response–stimulant concentration curve increases its slope as the stimulant reaches its population threshold, and then decreases its slope as it approaches saturation. In a model system, a short, steep response–stimulant concentration curve will correspond to an attribute that is difficult to scale, and translate to an on–off response for the panellist (Findlay, 2004). By contrast, a relatively slow rising response–stimulant curve will correspond to an attribute that is relatively easy to scale (Fig. 5).

Stimulant concentrations will rarely range from below population detection threshold to supersaturation for all sensory attributes within a real-world product group. In real-world complex products, sensory attributes will need to be present, often at similar levels, in order to be considered as a member of the product category. For example, the intensity of egg flavor may vary between mayonnaise products but will have a predictable minimum and maximum intensity, outside of which the product is no longer mayonnaise but cream sauce or whipped eggs. An attribute that is scalable within a model system may have a relatively narrow range of stimulant concentration in products being studied.

Products being evaluated may fall within a just noticeable difference, and may not be statistically discriminated, even by a well-trained descriptive panel using line scales. The absence of statistical discrimination from the panel could be considered meaningful information and not indicative of a panel's lack of proficiency. Comparing responses to training targets using the distance-from-range approach puts the question of proficiency in the context of whether the response is consistent or inconsistent with a response that would be expected from a well-trained descriptive sensory panel.

Data are analyzed by comparing observed responses (O_{ijk}) to expected responses (E_{ij}). This approach implies that all panellists are being encouraged to use the scale in the similar way. Monitoring a panel over successive training sessions using only distance from target measurements may, if the attributes intensities are similar across products and the training targets consistently within a particular region of the scale, provide a false impression of the panel's improvement; rapid decreases in distance from target in this case may not result from real increases in sensory acuity but simply from the panellists adjusting their scale usage patterns to use only a narrow location of the line scale to evaluate intensity. For this reason, conclusions that might be drawn from distance from target calculations should be validated using other statistical approaches, such as discrimination and disagreement.

A summary of the four distance from target measurements is presented in Table 1. Examples of comparisons that can be made using Distance from Target (DT) and Distance from Range (DR) measurements are presented together with their assumptions in Table 2.

5. Selected numerical applications

The study conducted by Findlay *et al.* (2003a, 2003b) included sensory data collected from two inexperienced panels, Panel E and Panel C, which differed in the method using to train the panels. Each panel consisted of 8 panellists. Both Panels E and C were trained to evaluate 20 red wines that had been previously evaluated by Panel D and for which training targets were available. All line scales were anchored at 0 and 100. Intermittent feedback, in the form of ellipses that represented the acceptable range, was provided to panellists for 31 line scale attributes for 10 of the 20 wines.

The training period consisted of 24 sessions, 14 of which were conducted in sensory booths. At each training session, 5 of the wines were selected and presented according to a modified 5-by-5 Williams' Latin square design. The focus of the experiment was on the training method and not on products; for this reason 4 evaluation days were interspersed with 10 training days on which feedback was presented, which for simplicity will be referred to as 1 to 10. The changes in panel calibration on the 10 training days were monitored using the |DT|... approach (Table 3). Results indicate similar improvement in the two panels during the study.

The relative success of the panellists can be monitored using frequency counts, as well as any of the distance from target approaches. Presented are frequency counts of in-range responses on a per panellist basis (Table 4), as well as |DR| for individual panellists (Table 5), both calculated from data collected on the last training session.

Results show there were varying degrees of calibration among the panellists on Panel E and panellists on Panel C. This sort of report could be useful when considering whether a panellist is improving, and to validate a decision to excuse a panellist from a standing trained panel.

The panel leader can also monitor a panel's or panellist's success in responding to individual attributes in a manner consistent with the training targets. This information could help a panel leader to determine whether the panel could benefit from reviewing particular attributes. For example, after the 6th training session, the panel leader of Panel E may decide to review the performance of the panel on all attributes and use this information to determine which attributes will be the focus of additional training.

Results of $|ADT|_j$ and $|ADT|_{.j.}$ are presented in Table 6, along with p_{wine} to indicate the panel's ability to discriminate the wines. Each p_{wine} value was obtained from a two-way mixed-model analysis of variance, with products treated as fixed effects and panellists as random effects.

When the panel is not well calibrated and is unable to discriminate wines for an attribute, the panel leader may conclude that more training is required. Table 6 suggests that the panel requires more training on attributes such as astringent and alcohol aroma and flavor. Conversely, the panel leader may be satisfied when both calibration and discrimination are excellent. The panel shows good calibration and discrimination on several attributes, including floral, rose, earthy/musty, cherry, and current aromas.

Honey aroma is an attribute on which Panel E shows good calibration ($|ADT|_{.\text{ honey aroma }.} = 2.9$) but poor honey aroma discrimination ($p_{wine} = 0.817$). Although Panel D successfully discriminated the 20 wines using the honey aroma attribute ($p_{wine} = 0.028$), it did not discriminate the 5 wines presented in the 6th session to Panel E using the honey aroma attribute, according to pairwise comparisons among samples using Tukey's HSD ($\alpha = 0.1$). Panel E did not discriminate the wines using honey aroma in the 6th session, but its responses are nonetheless considered to be consistent with those of a well trained panel.

There are cases in which a panel is able to discriminate wines in spite of relatively high calibration scores. Panel E discriminated wines using the pungent flavor attribute ($p_{wine} = 0.095$), in spite of problems with calibration on this attribute ($|ADT| = 444$). Panel D was able to discriminate the 20 pungent flavor wines using pungent flavor ($p_{wine} < 0.0001$), including the 5 wines presented to Panel E in the 6th session, according to Tukey's HSD ($\alpha = 0.1$). A breakdown of ADT by panellist for the pungent flavor attribute reveals that the largest contributor to |ADT| is Panellist E14 ($|ADT|_{\text{pungent flavor, E14}} = 206$), whose contribution to $|ADT|_{\text{pungent flavor}}$ is more than twice that of the next largest contributor, Panellist E01 ($|ADT|_{\text{pungent flavor, E01}} = 93.5$). Panellist E14 is using a different part of the scale to respond to pungent flavor. With this information, the panel leader can take immediate action to clarify attribute identity and scale usage for the pungent flavor attribute with this panellist.

$|ADR|_{...}$ can be calculated on a session-by-session basis for Panels E and C. The similarity in the trends of improvement of the panels can be shown by graphing $|ADR|_{...}$ as it changes during the 10 training sessions (Fig. 6). The two panels being compared were trained to perform descriptive analysis using different methods, but the graph reveals that the two panels are becoming calibrated to the training targets at similar rates.

$|ADT|_j$ refers to the summation of adjusted distance from target for all panellists and products for the *j*-th attribute. $|ADT|_{.j.}$ refers to the average adjusted distance from target for an observation for the *j*-th attribute. $p(wine)_j$ indicates the p-value of the wine effect for the *j*-th attribute calculated by submitting data for each attribute from Panel E for session 6 a two-way mixed-model ANOVA, where wine is treated as a fixed effect and panellist as a random effect.

6. Summary

Product profiles can provide an objective representation of the underlying sensory characteristics of a group of products. From these product profiles it is possible to extract training targets, which can be used to train new and maintain existing descriptive panels. Counting the number of responses for which the panel or individual panellists were in and out of range on a per-product, per-attribute, and overall basis can indicate the calibration of the panel. Additionally, panellist responses can be measured against these training targets using four approaches, Distance from Target, Distance from Range, Adjusted Distance from Target, and Adjusted Distance from Range, which are summarized in Table 6. Further investigation is required to explore how well Adjusted Distance from Target and Adjusted Distance from Range approximate the relative calibration of panels that are evaluating different products in different product categories using different lexicons.

Acknowledgements

The authors gratefully acknowledge financial support to Project #515979 and Project #474766 provided by Canada's National Research Council—Industrial Research Assistance Program (NRC-IRAP).

References

Findlay, C. J. (2004). Optimizing descriptive analysis (oral). In *The institute of food technologists annual meeting + food expo 2004*. Technical program: advances in sensory science symposium. 12–16 July, Las Vegas, NV.
Findlay, C. J., Castura, J. C., & Lesschaeve, I. (2003a). Feedback calibration: a training method for descriptive panels (oral). In *The fifth Pangborn sensory symposium*. 20–24 July, Boston, MA.
Findlay, C. J., Castura, J. C., & Lesschaeve, I. (2003b). Suspending continual feedback and its effect on panel performance (poster). In *The fifth Pangborn sensory symposium*. 20–24 July, Boston, MA.
Lawless, H. T., & Heymann, H. (1998). *Sensory evaluation of food, principles and practices*. Gaithersburg, MD: Aspen Publishers, Inc. (a Chapman & Hall Food Science Book).
McEwan, J. A. (2000). *Proficiency testing for sensory profile tests: statistical guidelines (part 1. R&D report no. 119)*. Chipping Campden, Gloucestershire, UK: Campden & Chorleywood Food Research Group.
Meilgaard, M., Civille, G. V., & Carr, B. T. (1999). Measuring response. In *Sensory evaluation techniques* (3rd ed., pp. 43–58). New York: CRC Press.

Table 1
Summary of distance from target measurements.

	Distance from Target (DT)	*Distance from Range (DR)*	*Adjusted Distance from Target (ADT)*	*Adjusted Distance from Range (ADR)*
Description	Responses measured against targets that are expected to reflect the most correct responses.	Responses measured against ranges that are expected to reflect the range in which responses are considered correct. Larger range sizes reflect relative difficulty in scaling attribute.	Reports DT as a proportion of the acceptable range.	Reports DR as a proportion of the acceptable range.
Assumptions	Accurate and stable training targets can be established for the product category.	The range size accurately reflects the difficulty that a well calibrated panellist will have in scaling the attribute for the product.	Range size accurately reflects the difficulty that a well calibrated panellist will have in scaling the attribute for the product.	Range size accurately reflects the difficulty that a well calibrated panellist will have in scaling the attribute for the product.
	See Table 2 for assumptions for particular comparisons.	See Table 2 for additional assumptions for particular comparisons.	If comparing different products or different attributes, range size can be used to adjust DT to allow for such comparisons.	If comparing different products or different attributes, range size can be used to adjust DR to allow for such comparisons.
Requirements	Targets	Ranges	Targets & Ranges	Ranges
Scale dependence	Yes Units as per scale	Yes Units as per scale	No Unitless measure	No Unitless measure

Table 2
Assumptions for some comparisons that can be made using Distance from Target (DT) and Distance from Range (DR) measurements.

Subject of comparison	*Distance from Target (DT)*	*Distance from Range (DR)*
panellist	Products, attributes, scales, and targets are the same among panellists being compared.	Products, attributes, scales, and ranges are the same among panellists being compared.
attribute	Scales and targets are the same among attributes being compared.	Scales and ranges are the same among attributes being compared.
*panellist*attribute*	Scales and targets are the same among panellists and attributes being compared.	Scales and ranges are the same among panellists and attributes being compared.
product	Attributes, scales, and targets are the same among products being compared.	Attributes, scales, and targets are the same among products being compared.
session	Products, attributes, scales, and targets are the same among sessions being compared.	Products, attributes, scales, and ranges are the same among sessions being compared.
overall	Products, attributes, scales, and targets are the same among panels being compared.	Products, attributes, scales, and ranges are the same among panels being compared.

Please refer to the published manuscript in *Food Quality and Preference* for all quotes and citations (doi:10.1016/j.foodqual.2005.03.011).

Table 3
Distance from target (|DT|...) across all panellists, products, and attributes on each of the last 10 training sessions.

Session	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Panel E	19.9	17.6	16.9	14.1	13.2	11.5	10.2	10.3	9.7	8.3
Panel C	18.9	19.0	14.7	13.0	12.5	10.9	9.5	8.6	8.3	8.3

Table 4
Overall in-range responses by panellist for all products and attributes on the last training session.

Panel C	*Observations in range*	*Percent in range*	*Panel E*	*Observations in range*	*Percent in range*
Panellist C02	90 / 155	58	Panellist E01	94 / 155	61
Panellist C02	101 / 155	65	Panellist E04	98 / 155	63
Panellist C05	82 / 155	53	Panellist E06	110 / 155	71
Panellist C07	62 / 155	40	Panellist E08	131 / 155	85
Panellist C10	110 / 155	71	Panellist E09	101 / 155	65
Panellist C12	77 / 155	50	Panellist E11	80 / 155	52
Panellist C13	140 / 155	90	Panellist E14	68 / 155	44
Panellist C16	83 / 155	54	Panellist E15	65 / 155	42
Overall	745 / 1240	60	Overall	747 / 1240	60

Please refer to the published manuscript in *Food Quality and Preference* for all quotes and citations
(doi:10.1016/j.foodqual.2005.03.011).

Table 5
Distance from range (|DR|) across all products and attributes on the last training session.

Panel C	$\lvert DR\rvert_{\cdot\cdot k}$	$\lvert ADR\rvert_{\cdot\cdot k}$	*Panel E*	$\lvert DR\rvert_{\cdot\cdot k}$	$\lvert ADR\rvert_{\cdot\cdot k}$
Panellist C02	2.62	0.39	Panellist E01	2.23	0.26
Panellist C02	2.38	0.28	Panellist E04	1.79	0.24
Panellist C05	3.87	0.47	Panellist E06	1.12	0.16
Panellist C07	6.69	0.98	Panellist E08	0.64	0.09
Panellist C10	0.88	0.12	Panellist E09	2.06	0.25
Panellist C12	3.54	0.49	Panellist E11	3.16	0.41
Panellist C13	1.58	0.16	Panellist E14	5.75	0.68
Panellist C16	3.60	0.45	Panellist E15	6.88	0.78
Overall	3.15	0.42	Overall	2.95	0.36

$\lvert DR\rvert_{\cdot\cdot k}$ indicates the per-observation DR for the *k*-th panellist. $\lvert ADR\rvert_{\cdot\cdot k}$ indicates the per-observation ADR for the *k*-th panellist. The Overall row for $\lvert DR\rvert_{\cdot\cdot k}$ and $\lvert ADR\rvert_{\cdot\cdot k}$ columns shows $\lvert DR\rvert_{\cdot\cdot\cdot}$ and $\lvert ADR\rvert_{\cdot\cdot\cdot}$, the per-observation DR and ADR, respectively, across all panellists, products, and attributes.

(doi:10.1016/j.foodqual.2005.03.011).

Table 6
The performance of Panel E on training session 6 as measured using adjusted distance from target.

		$\|ADT\|_{j}$	$\|ADT\|_{\cdot j \cdot}$	$p(wine)_j$
Aroma	Floral	165.5	4.1	0.062
	Rose	165.0	4.1	0.026
	Earthy/Musty	222.5	5.6	0.003
	Alcohol	477.5	11.9	0.199
	Pungent	394.5	9.9	0.533
	Oak Barrel	152.0	3.8	0.017
	Smoky	52.5	1.3	0.098
	Fermented	133.0	3.3	0.516
	Currant	174.0	4.4	0.029
	Cherry	144.5	3.6	0.024
	Sulphur	264.0	6.6	0.013
	Black Pepper	140.5	3.5	0.288
	Honey	115.0	2.9	0.817
	Medicinal	300.5	7.5	0.529
	Grape	274.5	6.9	0.169
	Red Berries	165.0	4.1	0.196
	Average	*208.8*	*5.2*	
Taste/Mouthfeel	Sweet	196.0	4.9	0.757
	Sour	303.0	7.6	0.025
	Bitter	325.5	8.1	0.345
	Salty	196.0	4.9	0.201
	Astringent	407.0	10.2	0.687
	Average	*285.5*	*7.5*	
Flavour	Earthy/Musty	174.0	4.4	0.934
	Oak Barrel	122.5	3.1	0.112
	Fermented	93.5	2.3	0.321
	Alcohol	438.0	11.0	0.275
	Pungent	444.0	11.1	0.095
	Yeasty	85.5	2.1	0.389
	Currant	228.5	5.7	0.608
	Cherry	153.5	3.8	0.442
	Grape	169.5	4.2	0.622
	Black Pepper	227.0	5.7	0.316
	Average	*213.6*	*5.3*	
Overall	***Average***	***222.7***	***5.6***	

$|ADT|_j$ refers to the summation of adjusted distance from target for all panellists and products for the *j* -th attribute. $|ADT|_{.j.}$ refers to the average adjusted distance from target for an observation for the *j*-th attribute. $p(wine)_j$ indicates the p-value of the wine effect for the *j*-th attribute calculated by submitting data for each attribute from Panel E for session 6 a two-way mixed-model ANOVA, where wine is treated as a fixed effect and panellist as a random effect.

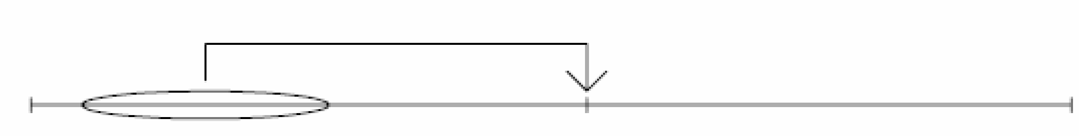

Fig. 1. Illustration of distance from target (DT). Distance from target is the differences between the panellist response, indicated by the mark under the arrowhead, and the discrete target, which usually occurs at the mid-point of the training target.

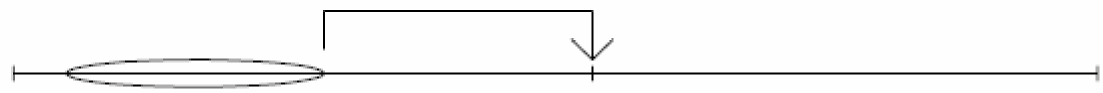

Fig. 2. Illustration of distance from range (DR). Distance from range is the differences between the panellist response, indicated by the mark under the arrowhead, and the nearest response that would fall within the training target. When a response falls within the training target, the distance from range is zero.

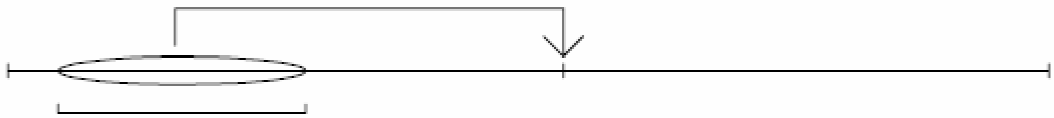

Fig. 3. Illustration of adjusted distance from target (ADT). Distance from target is divided by the size of the training target to produce the adjusted distance from target.

Please refer to the published manuscript in *Food Quality and Preference* for all quotes and citations (doi:10.1016/j.foodqual.2005.03.011).

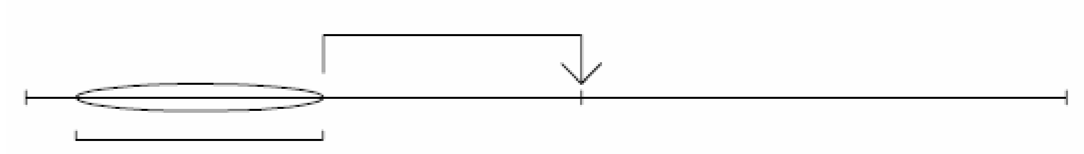

Fig. 4. Illustration of adjusted distance from range (ADR). Distance from range is divided by the size of the training target to produce the adjusted distance from range.

Please refer to the published manuscript in *Food Quality and Preference* for all quotes and citations (doi:10.1016/j.foodqual.2005.03.011).

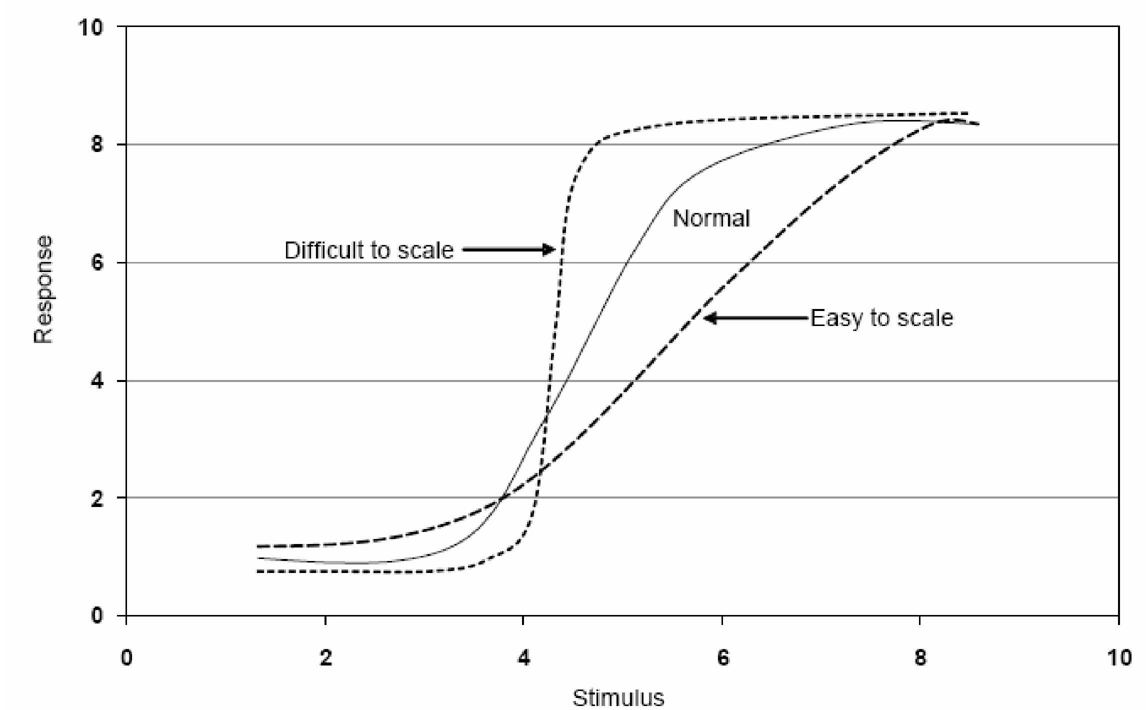

Fig. 5. Psychometric function showing stimulus–response plots for three attributes, one normal, one difficult to scale, and one easy to scale. The size of the training target will be smaller for the attribute that is easier to scale and larger for the attribute that is difficult to scale.

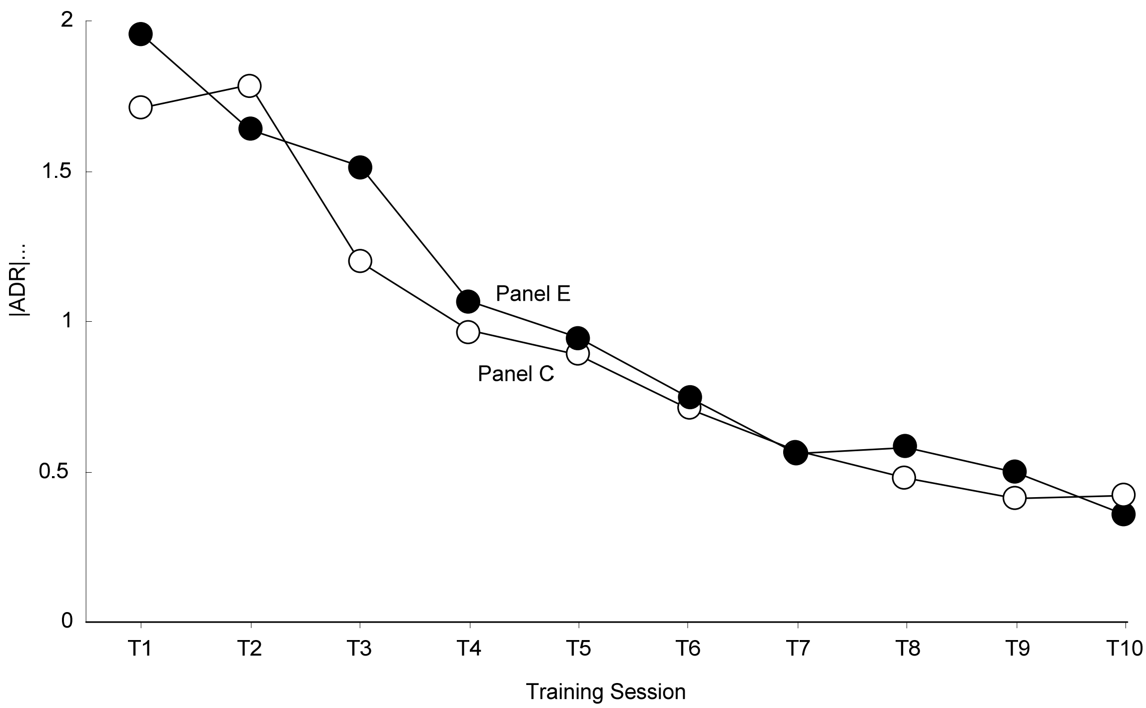

Fig. 6. Adjusted distance from range (|ADR|...) across all panellists, products, and attributes on each of the last 10 training sessions. The improvement of Panel E was comparable to that of Panel C during the training period.

Please refer to the published manuscript in *Food Quality and Preference* for all quotes and citations (doi:10.1016/j.foodqual.2005.03.011).