Generalized pair-wise correlation and method comparison: impact assessment for JAR attributes on overall liking

Attila Gere[1], László Sipos[1]*, Károly Héberger[2]

[1] Corvinus University of Budapest, Faculty of Food Science, Sensory Laboratory,
Villányi út 29-43, H-1118 Budapest, Hungary

[2] Research Centre for Natural Sciences, Hungarian Academy of Sciences,
Magyar Tudósok krt. 2, H-1117 Budapest, Hungary

* Corresponding author: E-mail: laszlo.sipos@uni-corvinus.hu

Abstract

In product development using JAR (just-about-right) scales, it is important to identify precisely, which direction of a given attribute affects hedonic scores the most. The Generalized Pairwise Correlation Method (GPCM) is a non-parametric one and it is useful to rank JAR variables according to their impact on liking. This is done using appropriate statistical tests: the McNemar's, the Chi-square, the Conditional Fisher's and the Williams' *t*-test. As GPCM requires one-directional variables, JAR data needs to be transformed based on the dummy variable approach. GPCM gives those attributes in that order, which should be increased/decreased to gain higher consumer liking scores. An order can be created according to the impact on liking, which order determines the development of product attributes, as well. The non-parametric tests incorporated in the method are able to identify smaller differences than other statistical methods. As a result, GPCM identifies more significant product attributes; hence, it can help product development processes even if other methods cannot.

Keywords: Generalized Pairwise Correlation, just-about-right data analysis, Penalty Analysis, product optimization, feature selection, software

1. Introduction

Just-About-Right (JAR) scales are bipolar scales used to measure the level of an attribute relative to a sensory assessor's ideal level, having a midpoint labeled "just about right" or "just right". A scale is bipolar when the end anchors are semantic opposites; for example, "not nearly sweet enough" to "much too sweet," and there is an implied or anchored neutral mid-point (ASTM E253; ASTM E456). A discussion about the meaning of the JAR levels was published by Gacula *et al*. in 2007.

Figure 1

The number of categories used in the JAR scales usually varies between three and nine. Figure 1 shows a JAR scale consisting of five categories. In consumer research and sensory sciences, JAR scales are typically measured along with hedonic scores: the relationship between the JAR variable and the hedonic score is used for either optimization or further development of products. JAR scales have been used to optimize salsas (Popper & Gibes, 2004), raisin jams (Rababah *et al*., 2012), probiotic Petit Suisse Cheese (Esmerino *et al*., 2013), mixed juices from Amazon fruits (Freitas & Mattietto, 2013), mango nectar (Cadena & Bolini, 2012), cooked steaks (Chan *et al*., 2013), Bulgogi (Korean traditional barbecued beef) (Hong *et al*., 2011) and juice blends (Lawless *et al*., 2013), just to name a few. Despite frequent usage, there are several unsolved issues regarding the JAR scales: (i) consumers must understand the attributes in question; (ii) the endpoints must be true opposites (Lawless & Heymann, 2010); (iii) there is no conclusive evidence that the use of JAR scales predicts optimal products (Epler *et al*., 1998). Earthy *et al*. (1997) and Popper *et al*. (2004) showed that JAR questions present in the questionnaire have an effect on overall liking and some authors do not recommend the use of JAR scales (Stone & Sidel, 2004). On the other hand, others report that JAR methods perform well during product development as we mentioned above; (iv) it is not known how much of products should be adjusted to reach optimal JAR likeness. Lesniauskas and Carr concluded that JAR data analysis should answer three questions (Popper & Gibes, 2004):

(1) Are some products more JAR than others?

(2) If the product is not JAR, in which direction should it be advanced?

(3) How much is the acceptance affected when a product is not JAR?

Several methods have been introduced to evaluate JAR data: Penalty Analysis (ASTM MNL63), Multivariate Adaptive Regression Splines (MARS) (Xiong & Meullenet, 2004), Partial Least Squares Regression (PLS-R) with dummy variables (Xiong & Meullenet, 2006), Canonical Variate Analysis (CVA) (Popper & Gibes, 2004) and Thurstonian Ideal Point (TIP) modeling (Goerlitz & Delwiche, 2004).

Penalty Analysis appears to be the most commonly used method to analyze JAR data. Penalty Analysis consists of three main steps in case of a five point JAR scale. Firstly, the JAR values are amalgamated into three groups. Categories 1 and 2, category 3, and category 4 and 5 give the three new levels: "not enough", "JAR", and "too much". Then, the mean overall liking (rating) is calculated for each group. The penalties (or mean drops) are calculated as the differences between the means of the two non-JAR categories and the mean of the JAR category. These values are plotted *versus* the percentage giving each response in a so-called mean drop plot.

Several papers dealing with the improvement of the method have been published in the past few years. Plaehn and Horne (2008) introduced a regression-based approach for testing the significance of JAR variable penalties. Plaehn (2013) created a Penalty Allocation Map (PAM) to improve the estimation of the “penalties”. Pagès *et al*. (2014) used Multiple Correspondence Analysis (MCA) to visualize the uncertainty in penalties. The main aim of analyzing JAR data is to identify the strengths/weaknesses of a given product and to determine, which attributes should be increased or decreased in future product formulations.

Most of the above mentioned methods are able to reach this goal, but for product developers it would be very useful to identify precisely, which direction (not enough or too much) of a given attribute affects the hedonic scores the most. Feature selection methodologies (methods applied to define a subset of relevant variables in model construction) would answer this question after proper data pretreatment.

A nonparametric approach for feature selection was introduced by Rajkó and Héberger, named Pairwise Correlation Method (PCM), which can discriminate between two features (Rajkó & Héberger, 2001). Let us assume that a dependent feature (***Y***) contains the hedonic scores and that two independent features (JAR variables, ***X***1 and ***X***2) have been measured. PCM makes a distinction between the two features. This idea was later generalized (Generalized Pairwise Correlation Method, GPCM) up to several hundred features (Héberger & Rajkó, 2002a, 2002b).

The present paper introduces this new technique, which uses a different set of information present in the data than usual. The aim of this study is to investigate the usefulness of GPCM to identify the main factors influencing the hedonic scores. The outcome helps to create proper regression models for predicting hedonic scores or building preference maps.

2. Materials and Methods

2.1 Data analysis

2. 1.1. Coding of the JAR data

The bipolar Just-About-Right (JAR) scales cannot be evaluated using linear approaches, as the optimal point is located above the points next to it (Figure 2). The JAR data set consists of ratings from the consumers. These ratings are not normally distributed and furthermore the scale has two directions. In the case of a general five point JAR scale, the midpoint (rating 3) belongs to the term “just about right”, rating 1 belongs to the term “the attribute is too weak” and rating 5 belongs to the term “the attribute is too strong”. Therefore, two linear relations would be suitable by just cutting the JAR scale in the middle in two parts. In creating a one directional scale, loss of information is inevitable because the number of consumers, who rated the product as JAR, cannot be calculated.

Figure 2

The bipolar scale can be transformed into a unipolar one by applying the dummy variable approach introduced by Xiong and Meullenet (2006). This transformation has to be performed before running Generalized Pairwise Correlation Method (GPCM). A possible creation of one directional scales is shown in Table 1.

Table 1

2.1.2 Pairwise Correlation Method (PCM)

PCM is a non-parametric method, developed for the discrimination of features. It has the advantage of non-parametric statistics so the assumption of normality is not required. Let us define three vectors as a dependent variable (***Y*** or consumer’s liking) and two independent variables ($\boldsymbol{X}1$ and $\boldsymbol{X}2$ or JAR attributes). Both of the $\boldsymbol{X}1$ and $\boldsymbol{X}2$ features have positive correlations with the ***Y*** feature. The task is to choose for the response feature (***Y***) the superior one from the coequal predictor features $\boldsymbol{X}1$ and $\boldsymbol{X}2$. Consider all the possible element pairs for $\boldsymbol{Y}_i$ *vs.* $\boldsymbol{X}1_i$ and $\boldsymbol{X}2_i$ as well as $\boldsymbol{Y}_j$ *vs.* $\boldsymbol{X}1_j$ and $\boldsymbol{X}2_j$ (running coordinates are *i* and *j* 1, 2, … *N*) (see Figure 3). Four basic events can be counted *i.e. A, B, C,* and *D*. The frequencies of the events *A, B, C,* and *D* (k_A, k_B, k_C and k_D, respectively) are counted and ordered (see Table 2). Event *A* means that both ***X***1 and ***X***2 strengthen the changes in ***Y***. Event *B* means that ***X***1 strengthens but ***X***2 weakens, while event *C* means that ***X***2 strengthens but ***X***1 weakens the same changes. Event *D* means that both ***X***1 and ***X***2 weaken the changes in ***Y*** (Figure 3) (Rajkó & Héberger, 2001).

Table 2

Figure 3

The frequencies for the four possible events are arranged in a 2×2 contingency table. If both differences (Δ***X***1 and Δ***X***2) are positive (or both are negative), the distinction cannot be made between ***X***1 and ***X***2. If the frequency value for opposite signs of Δ***X***1 is significantly greater, then ***X***1 is termed as superior, otherwise ***X***2. Whether the frequency value is significantly greater (or not) can be determined by the next statistical tests: the McNemar's, the Chi-square and the Conditional Fisher's tests as non-parametric statistical ones. For the sake of comparison, the Williams' *t*-test as a parametric test was also included.

The program calculates the critical sum and the limit (theoretical) probability at which the frequencies in *B* and *C* boxes can still be considered as significant. An example shows a case of clear superiority of a feature: Table 3 contains the comparison of Color+ and Carbonation+ where Carbonation+ overrides Color+ i.e. (wins) unambiguously.

Table 3

In its generalized form (GPCM), all possible independent feature pairs are compared and a number of "superiority" is determined. The number of "superiority" is termed as the number of wins, *i.e*. how many times a given ***X*** feature was "superior" to the other ***X*** features. The number of "inferiority" is termed as the number of losses; *i.e*. how many times an ***X*** feature was inferior to the other ***X*** features. The number of wins is simply summed for all possible comparisons of feature pairs. Three ranking methods (a) simple ranking according to the number of wins, (b) ranking according to the differences in wins and losses and (c) probability weighted ranking according to the differences in wins and losses were elaborated in the Generalized Pairwise Correlation method (Héberger & Rajkó, 2002a).

It has to be mentioned that GPCM might be sensitive to outliers in ***Y*** direction (dependent variable) either on the plot of ***X***1 variable or in the plot of the ***X***2 variable. Such outliers of ***Y*** influence the frequencies of events *B* or *C*. If a ***Y*** outlier is present in both plots of ***X***1 and ***X***2, then, the frequency of event *D* is influenced and would not be evaluated further. Random fluctuations do not affect the results of GPCM. The method uses systematic information; hence outliers have to be removed before analysis. Let us notice that logical operations require more calculation time than arithmetical operations, hence the high number of observations increase computation time, but this is not mandatory.

2.1.3 Generalization of PCM based on simple ranking

Only the number of wins is counted in the case of simple ranking; *i.e.* the features are ranked according to the number of wins. The features, which did not win at all, are excluded automatically. A percentage (*e.g.* 95 %) of the total number of wins indicates the number of variables, which have to be selected (Héberger & Rajkó, 2002a).

2.1.4 Generalization of PCM based on ranking according to differences

PCM has to be carried out pairwise for all possible (different) feature pairs as above. In case of this GPCM ranking, the features are ranked according to the number of wins minus the number of losses. A percentage value (*e.g.* 95 %) in the number of wins minus number of losses indicates the number of features which have to be selected. Features, which did not win at all and those for which the number of wins is less than or equal to the number of losses are not selected to be significant features. However, the numbers of wins minus losses are suitable to rank all of the features (just the zero and negative values should not be excluded). This ordering method is the second most sensitive but data characteristics can also influence the number of significant differences.

2.1.5 Generalization of PCM based on ranking according to probability-weighted differences

PCM has to be done for all possible different pairs as in the previous sections. Wins and losses are counted as above, but they are weighted with the calculated confidence level (=1-calculated error level) based on any test statistics. In this way the features are ranked according to the sum of the weighted differences. Subtracting the error levels, the numbers of wins and losses will not be integer numbers (Héberger & Rajkó, 2002a). This type of ranking produces the most sensitive one, it takes into account whether the wins were overwhelming or not.

GPCM data analysis was done using a Microsoft Excel VBA spreadsheet application:

http://aki.ttk.mta.hu/gpcm

Spearman’s rank correlation, Multiple Linear Regression (MLR), Partial Least Squares Regression (PLS-R) and Penalty Analysis was done using XL-Stat Sensory solution (Addinsoft, 28 West 27th Street, Suite 503, New York, NY 10001, USA).

2.2 JAR experiment

2.2.1 Materials

In our study, flavored mineral water samples were evaluated by consumers. The tested samples were prototypes with the mango-passion fruit aroma.

2.2.2 Sample preparation

The preparation of the samples (stored at 10 °C) was conducted using the same standardized parameters (refrigerator, sample quantity, material and brand of the glasses, etc.). Samples were poured into plastic glasses. The recommendations of Kilcast were followed during the sample presentation, so the quantities of samples (180 cm^3/person) were prepared by one person using a measuring cup to achieve better homogeneity (Kilcast, 2010). Samples were labeled, according to the international practice using 3-digit random numbers and a balanced block design was applied (ISO 6658:2005). The samples were presented to the assessors in plastic glasses (200 cm^3) at a typical consumption temperature (15 °C), which was strictly monitored to maintain commensurable conditions. Between the evaluations of the products, assessors used a very neutral non-carbonated mineral water as taste neutralizer (Sipos *et al.*, 2012). Evaluations were performed under artificial daylight-type illumination, temperature control (between 22 and 24 °C) and air circulation.

2.2.3 Consumer test

Næs *et al.* suggests a minimum of 60 consumers for these kinds of tests (Næs *et al.*, 2010); hence 117 consumers were recruited from the Corvinus University of Budapest, Hungary. Consumers were selected according to relevant market figures: 60 % / 40 % males/females aged between 18 and 30 years, regular carbonated soft drink (CSD) consumers, as they consumed CSD products more than 2-3 times a week. Consumers were instructed prior to the evaluation to ensure the reliability of the results and asked to evaluate overall liking on a 9-point hedonic scale (1 = ‘‘dislike extremely”, 9 = ‘‘like extremely”). The attributes of color intensity, odor intensity, fruit flavor, carbonation, sweet taste, sour taste, bitter taste and aftertaste intensity were evaluated using a five point ‘‘Just About Right” scale (Figure 1). In this study, the data set of the most preferred product will be presented.

3. Results

Table 4 lists the most important product attributes. Based on the ordering methods, simple ordering has resulted in the most sensory attributes (out of the 16) as significant. The non-parametric tests (Conditional exact Fisher's test (CondExact), McNemars, ChiSquare) show significant effects on the liking in case of at least ten attributes.

Table 4

The last column in Table 4 illustrates the marginal efficiency of a parametric test: only three attributes were significant, of which the first two were identical to the results given by the non-parametric tests. Simple ordering method has produced the largest number of significant variables, because this is the least strict method. In contrast, difference ordering and significance ordering have more and stricter conditions (see Sections 2.1.3-2.15).

The difference ordering method, which takes into account the number of losses, has a similar result as that of the significance ordering. The main difference to the simple ordering is the number of significant attributes, which were reduced to 2-5 depending on the applied test.

Significance ordering is the most strict and least sensitive ordering method. The ranks of the attributes do not change except in the case of the McNemar's test. In this case the too weak sweet taste is on the fourth rank, but in the case of simple ordering this attribute goes to the fifth rank (Table 4). As it is worth to consider the consensus of ordering methods, the three most important attributes were FruitFlav-, Aftertaste+ and Bittertaste+, in decreasing order.

Based on these results, the main direction of product development should be increasing the fruit flavor. However, reduced aftertaste and bitter taste intensity would also increase the liking of consumers.

A new visualization tool summarizes the results of the three ordering methods and three statistical tests (Figure 4). The first rank is given to the attribute which has the highest influence on consumer's liking (FruitFlav-). The second rank is given to the second most influential (Aftertaste+), and the third is to the third most influential (Bittertaste+) and so on. Horizontal lines represent the consensus; crossing lines represent the lack of consensus between the tests and ordering methods. Connection of the points is done to enhance interpretability of the plot. It is easier to compare the ordering methods, if they are grouped next to each other on the *x*-axis.

Figure 4

Another new, Penalty Analysis-like visualization gives the percentage of the consumers *vs*. the ranks given by the GPCM method using bubble plot (Figure 5).

Figure 5

The plot is divided into four subspaces using two lines. The horizontal line represents 20 % of the consumers, while the vertical line represents the border of the significant attributes. The size of the bubbles defines the rank of the attributes. The upper left subspace contains the significant and important (> 20 % of consumers' rates) attributes. These attributes should be emphasized during product development. The bottom left subspace contains the significant but not important (< 20 % of consumers' rates) attributes. The upper right quadrant shows the non-significant but important attributes. These attributes have little effect on consumers liking but more than 20 % of the consumers rated as true. The bottom right subspace contains the non-significant and not important attributes. The bigger the size of a bubble, the higher its impact on liking.

The rank of the attributes is based on their impact on consumers liking (*x* coordinate and the size of the bubble), while the *y* coordinate gives the percentage of the consumers. For example, in the case of color, 30 % of the consumers said that it is not intensive enough (Color-), but it has no significant effect on liking (its rank is 16). In contrast, 5 % of the consumers rated the color as too intensive (Color+), but for these consumers this had significant effect on their liking.

To validate the results of GPCM, the following methods were used: Spearman's rank correlation, Multiple Linear Regression (MLR) (ASTM MNL63), Partial Least Squares Regression (PLS-R) (Xiong & Meullenet, 2006) and Penalty Analysis (ASTM MNL63).

GPCM does not require high correlations between the independent and dependent variables. Due to the non-parametric nature of JAR variables, the relationships between the attributes and liking were defined using Spearman's rank correlation coefficients (Table 5).

Table 5

Out of the 16 investigated variables, eight had significant correlation with the dependent variable (α=0.05): FruitFlav-, Aftertaste+, Sweettaste-, Bittertaste+, Odor-, SourTaste-, Carbonation+ and

Sourtaste+, respectively. The Spearman's rank correlation coefficients support and validate the results of GPCM. The rank of the attributes created by both methods is identical, c.f. Table 4.

Using MLR, out of the 16 attributes, seven had significant positive or negative effect on overall liking: FruitFlav- (positive effect), Aftertaste+ (negative effect), Carbonation+ (negative effect), FruitFlav+ (negative effect), Aftertaste- (positive effect), Carbonation- (positive effect), Sourtaste+ (negative effect). Thus, these attributes should be addressed. The sign (positive or negative) suggests how to reformulate the product. If the sign is positive, more is better (up to a certain point). Likewise, if the sign is negative, less is better. FruitFlav+ and Aftertaste+ proved to be the first two most important attribute based on the *t*-values. Table 6 shows the results of MLR; the attributes are sorted according to their *t*-value.

Table 6

PLS-R was also used to identify variables, which have significant effect on liking. Compared with MLR, our PLS-R model had a slightly lower R^2 (0.480 *vs*. 0.508), and RMSE value (1.347 *vs*. 1.360). MLR had lower intercept (7.837), while our PLS-R model gave a value of 7.673, which can be interpreted as the estimated mean of overall liking for the product if all JAR attributes are JAR. Table 7 displays the Variable Importance for the Projection (VIP) values for each explanatory feature that measure the importance of each feature. This allows the quick identification of those explanatory features that contribute the most to the model.

Table 7

The VIP results of our PLS-R model showed that FruitFlav- (positive effect), Aftertaste+ (negative effect), Sweettaste- (positive effect), Odor- (positive effect), Bittertaste+ (negative effect), Sourtaste- (positive effect), Carbonation+ (negative effect) and Carbonation- (positive effect) have significant effect on liking.

On the VIP chart (Figure 6) (one bar per component), a border line is plotted to identify the VIPs, which are greater than 0.8. These thresholds, suggested by Wold (1995), allow identifying the features which are moderately (0.8<VIP<1) or highly influential (VIP>1).

Figure 6

Penalty Analysis identified the following attributes as significant ones (sorting is based on the mean drop values): FruitFlav-, Aftertaste+, Sweettaste-, Sweettaste+, Carbonation+, Sourtaste-, Carbonation-, Aftertaste-, Odor-. According to the mean drop plot (Figure 7), FruitFlav- seems to be the most important but the importance of Aftertaste+ and Sweettaste- is not highlighted as much as GPCM does.

In Penalty Analysis, the generally applied one sample *t*-test is used to compare the JAR and the two other endpoints and the test requires normally distributed variables. The results of Shapiro-Wilk's tests show, which variables do not fulfill the criteria of normality. In Table 8, Shapiro-Wilk's test *p*-values in bold indicate the non-violation of normality. As Table 8 indicates, the *t*-test was used several times on not normally distributed data.

Table 8

Figure 7

Table 9

It can be seen in Table 9 that the sequence of the first two significant variables is identical in the case of all JAR methods. The order of variables varies according to the differences of the applied methods and criteria. The highest number of variables were found in the case of GPCM simple ordering CondExact method and the strictest were GPCM difference ordering and significance ordering methods. Though the rankings of the attributes were similar; some switches were found in some cases.

4. Discussion and Conclusions

Generalized Pairwise Correlation Method (GPCM) proved to be suitable for analyzing the impact of Just-About-Right (JAR) variables on liking. An order can be created according to the impact on liking, in which order determines the development of product attributes as well. The non-parametric tests incorporated in the method are able to identify small differences and they have proved to be more sensitive than other statistical methods (*e.g.*: Williams' *t*-test). GPCM can provide differences when other statistical tests cannot. GPCM identifies more significant product attributes, so it can help the product development process.

One of the advantages of GPCM comes from its non-parametric nature. Hence, it is better suited to JAR variables which generally do not follow normal distribution and have significant correlations with each other. GPCM chooses relevant independent variables, which means higher safety when interpreting the results. Moreover, it is based on well-known and well-proven statistical criteria. GPCM requires one-directional scales, but they do not cause considerable loss of information and they preserve the dominant information of JAR scales (Xiong & Meullenet, 2006).

However, the proposed GPCM method cannot be used for prediction in a way similarly to Penalty Analysis. Besides, due to the characteristics of the method, the result is always an ordered list of variables. It will provide a result even if the first variable has a very low impact on the dependent variable. In contrast, GPCM is able to rank seemingly undistinguishable variables as well.

A new visualization tool (line plot) summarizes the results of the three ordering methods and three statistical tests. Of the three ranking methods, the simple ranking is the least conservative (it selects more descriptors). The best method, which can utilize even small differences in statistical tests, is the ranking according to probability-weighted differences. In our case study the ranks given by difference and significance ordering were similar. Based on this, the suggested order of product development should be the following: FruitFlav-, Aftertaste+, Bittertaste+, Sweettaste-, Sourtaste+. Another new, easily interpretable visualization tool was proposed to show the significant and important (> 20 % of consumers' rates) attributes in one single bubble plot.

Results of Spearman's rank correlations support and validate the results of GPCM. It is noteworthy that in our case study both approaches led to similar conclusions even if the train of thought is totally different. To validate a more precise method with accepted less precise ones (GPCM with simple ordering and Spearman rho; non parametric tests and Williams *t* test) is virtually impossible. Spearman rho shows that the results are in conformity (not in contradiction) with the earlier, well-known methods. Again, the train of thought of GPCM is totally different from the Spearman rho.

Multiple Linear Regression (MLR) has some model restrictions: *e.g.* number of cases ≥ number of variables, overfitting, signs of the regression coefficients cannot be interpreted in many cases, and strong correlation between the variables reduces the prediction ability (Krishnamurthy *et al.*, 2007). In contrast, Partial Least Squares Regression (PLS-R) handles the strongly correlated variables. Both PLS-R and MLR use the information present in the dependent variable, but PLS-R uses the information present in ***Y*** (Næs *et al.*, 2010) in a better way.

The main advantage of GPCM, compared to Penalty Analysis, is that it applies three ordering methods and four statistical tests to determine the sequence of variables having the greatest impact on liking. Penalty Analysis does not take into account the correlation between variables

(JAR variables tend to have strong correlations) and cannot be used to predict liking data (Plaehn, 2013).

Using GPCM, MLR, PLS-R and Penalty Analysis, the two most important attributes, which have high impact on liking, were determined unambiguously: FruitFlav- and Aftertaste+. The specificity of MLR and PLS-R is that they have been developed primarily for prediction purposes. But in the case of JAR data analysis the main aims are ranking and selecting attributes, determining the sequence of the significant attributes but not the prediction.

Acknowledgement

The authors would like to thank Robert Atkins for his helpful assistance in proof reading. This work was supported by the research project of the National Development Agency TÁMOP 4.2.1.B-09/1/KMR-2010-005. KH thanks the OTKA support, contract No K112547.

References

ASTM E253 Terminology Relating to Sensory Evaluation of Materials and Products, Conshohocken: ASTM International

ASTM E456 Terminology Relating to Quality and Statistics, Conshohocken: ASTM International

ASTM MNL63 Just-About-Right (JAR) Scales: Design, Usage, Benefits, and Risks, Conshohocken: ASTM International

Cadena, R. S., & Bolini, H. M. A. (2012). Ideal and relative sweetness of high intensity sweeteners in mango nectar. *International Journal of Food Science and Technology, 47*(5), 991-996.

Chan, S. H., Moss, B. W., Farmer, L. J., Gordon, A., & Cuskelly, G. J. (2013). Comparison of consumer perception and acceptability for steaks cooked to different endpoints: Validation of photographic approach. *Food Chemistry, 136*(3-4), 1597-1602.

Earthy, P.J., Macfie, H.J.H. & Hedderley, D. (1997). Effect of question order on sensory perception and preference in central location trials. *Journal of Sensory Studies 12*, 215–237.

Epler, S., Chambers, E., & Kemp, K. E. (1998). Hedonic scales are a better predictor than just-about-right scales of optimal sweetness in lemonade. *Journal of Sensory Studies, 13*(2), 191-197.

Esmerino, E. A., Cruz, A. G., Pereira, E. P. R., Rodrigues, J. B., Faria, J. A. F., & Bolini, H. M. A. (2013). The influence of sweeteners in probiotic Petit Suisse cheese in concentrations equivalent to that of sucrose. *Journal of Dairy Science, 96*(9), 5512-5521.

Freitas, D. D. C., & Mattietto, R. D. (2013). Ideal sweetness of mixed juices from Amazon fruits. *Food Science and Technology, 33*, 148-154.

Gacula, M., Rutenbeck, S., Pollack, L., Resurreccion, A. V.A. & Moskowitz, H. R. (2007). The just-about-right intensity scale. Functional analyses and relation to hedonics. *Journal of Sensory Studies, 22*, 194–211.

Goerlitz, C. D., & Delwiche, J. F. (2004). Impact of label information on consumer assessment of soy-enhanced tomato juice. *Journal of Food Science, 69*(9), S376-S379.

Heberger, K., & Rajko, R. (2002a). Generalization of pair correlation method (PCM) for nonparametric variable selection. *Journal of Chemometrics, 16*(8-10), 436-443.

Heberger, K., & Rajko, R. (2002b). Variable selection using pair-correlation method. Environmental applications. *SAR and QSAR in Environmental Research, 13*(5), 541-554.

Hong, J. H., Yoon, E. K., Chung, S. J., Chung, L., Cha, S. M., O'Mahony, M. (2011). Sensory Characteristics and Cross-Cultural Consumer Acceptability of Bulgogi (Korean Traditional Barbecued Beef). *Journal of Food Science, 76*(5), S306-S313.

ISO: Sensory analysis — Methodology — General guidance. International Standard 6658:2005, 2005.

Kilcast, D. (2010). Sensory quality control for taint prevention. In Kilcast D (ed.), *Sensory analysis for food and beverage quality control* (pp. 156-185). Cambridge: Woodhead.

Krishnamurthy, R., Srivastava, A. K., Paton, J. E., Bell, G. A., & Levy, D. C. (2007). Prediction of consumer liking from trained sensory panel information: Evaluation of neural networks. *Food Quality and Preference, 18*(2), 275-285.

Lawless, H. T. & Heymann, H. (2010). *Sensory Evaluation of Food: Principles and Practices.* (2nd ed.). New York: Springer, (Chapter 10).

Lawless, L. J. R., Threlfall, R. T., Meullenet, J. F., & Howard, L. R. (2013). Applying a Mixture Design for Consumer Optimization of Black Cherry, Concord Grape and Pomegranate Juice Blends. *Journal of Sensory Studies, 28*(2), 102-112.

Næs, T., Brockhoff, P. B., & Tomic, O. (2010). *Statistics for sensory and consumer science.* Chichester: John Wiley and Sons Ltd, (pp. 11–34, pp. 127-224).

Pages, J., Berthelo, S., Brossier, M., & Gourret, D. (2014). Statistical penalty analysis. *Food Quality and Preference, 32*, 16-23.

Plaehn, D. (2013). What's the real penalty in penalty analysis? *Food Quality and Preference, 28*(2), 456-469.

Plaehn, D., & Horne, J. (2008). A regression-based approach for testing significance of "just-about-right" variable penalties. *Food Quality and Preference, 19*(1), 121-32.

Popper, R., & Gibes, K. (2004). Workshop summary: Data analysis workshop: getting the most out of just-about-right data - Abstracts. *Food Quality and Preference, 15*(7-8), 891-899.

Popper, R., Rosenstock, W., Schraidt, M. & Kroll, B.J. (2004). The effect of attribute questions on overall liking ratings. *Food Quality and Preference*. *15*, 853–858.

Rababah, T. M., Al-u'datt, M., Almajwal, A., Brewer, S., Feng, H., Al-Mahasneh, M. (2012). Evaluation of the Nutraceutical, Physiochemical and Sensory Properties of Raisin Jam. *Journal of Food Science, 77*(6), C609-C613.

Rajko, R., & Heberger, K. (2001). Conditional Fisher's exact test as a selection criterion for pair-correlation method. Type I and Type II errors. *Chemometrics and Intelligent Laboratory Systems, 57*(1), 1-14.

Sipos, L., Kovacs, Z., Sagi-Kiss, V., Csiki, T., Kokai, Z., Fekete, A. (2012). Discrimination of mineral waters by electronic tongue, sensory evaluation and chemical analysis. *Food Chemistry, 135*(4), 2947-2953.

Stone, H., & Sidel, J. (2004). *Sensory Evaluation Practices*, New York: Elsevier Academic Press, (p. 92).

Wold, S. (1995). PLS for multivariate linear modelling. In: van de Waterbeemd H. (ed.), *QSAR: Chemometric Methods in Molecular Design. Vol. 2* (pp. 195-218). Weinheim: Wiley-VCH.

Xiong, R., & Meullenet, J. F. (2004). Application of Multivariate Adaptive Regression Splines (MARS) to the preference mapping of cheese sticks. *Journal of Food Science, 69*(4), S131-S139.

Xiong, R., & Meullenet, J. F. (2006). A PLS dummy variable approach to assess the impact of jar attributes on liking. *Food Quality and Preference, 17*(3-4), 188-198.

Table 1

Scheme for using two dummy variables (Aftertaste+ and Aftertaste-) or one dummy variable (Aftertaste) to represent one JAR scale variable (X) on a 5-point JAR scale based on Xiong and Meullenet (2006). All the attributes are indicated based on this table.

Original JAR values (5 point JAR scale)	Aftertaste	Aftertaste+	Aftertaste-
1 – much too weak	-2	-2	0
2 – somewhat too weak	-1	-1	0
3 – just about right	0	0	0
4 – somewhat too strong	1	0	1
5 – much too strong	2	0	2

Table 2

Frequencies obtained by applying GPCM

	$\Delta \boldsymbol{X}_1>0$	$\Delta \boldsymbol{X}_1<0$
$\Delta \boldsymbol{X}_2>0$	$k_A =$	$k_C =$
$\Delta \boldsymbol{X}_2<0$	$k_B =$	$k_D =$

Table 3

Comparison of Color+ and Carbonation+ features based on Conditional Fisher's exact test, The contingency table contains frequencies:

	Delta(Color+)>0	Delta(Color+)<0
Delta(Carbonation+)<0	$k_D = 58$	$k_B = 9$
Delta(Carbonation+)>0	$k_C = 64$	$k_A = 89$

Naturally, the equal $\boldsymbol{Y}$ values do not carry any information, they were ignored; critical value = 30, Predefined error limit α(user) = 0.05, Theoretical (threshold) α= 0.0000; Carbonation+ won, as nine is (much) less than 30. Delta means the difference in JAR attributes as described in 2.1.2 part.

Table 4

Rank of the attributes of product mango-passion fruit (all GPCM methods and all applied tests)

	Attributes	**CondExact**	**McNemars**	**ChiSquare**	**Williams' *t***
Simple ordering	*FruitFlav-*	1	1	1	1
	Aftertaste+	2	2	2	2
	Bittertaste+	3	3	3	NS
	Sweettaste-	4	5	5	3
	Sourtaste+	5	4	4	NS
	Carbonation+	6	6	6	NS
	Odor-	7	7	7	NS
	Sourtaste-	8	8	8	NS
	Sweettaste+	9	10	10	NS
	Color+	10	9	9	NS
	Carbonation-	11	NS	11	NS
	Odor+	NS	NS	12	NS
Difference ordering	*FruitFlav-*	1	1	1	1
	Aftertaste+	2	2	2	2
	Bittertaste+	3	3	3	NS
	Sweettaste-	4	NS	5	NS
	Sourtaste+	5	NS	4	NS
Significance ordering	*FruitFlav-*	1	1	1	1
	Aftertaste+	2	2	2	2
	Bittertaste+	3	3	3	NS
	Sweettaste-	4	4	5	NS
	Sourtaste+	5	NS	4	NS

Attributes in bold indicate the “too much” region, attributes in italic indicate the “too weak” region of the JAR scale. NS stands for non significant.

Table 5

Spearman's rank correlation matrix of the evaluated attributes. Attributes are ordered according to their correlation coefficients.

Features[a]	Overall liking[b]
FruitFlav-	**0.477**
Aftertaste+	**-0.342**
Sweettaste-	**0.315**
Bittertaste+	**-0.305**
Odor-	**0.227**
Sourtaste-	**0.224**
Carbonation+	**-0.213**
Sourtaste+	**-0.188**
Carbonation-	0.155
Sweettaste+	-0.155
FruitFlav+	-0.082
Aftertaste-	0.074
Odor+	0.07
Color+	0.065
Bittertaste-	-0.012
Color-	0.001

[a] Attributes in bold indicate the "too much" region, attributes in italic indicate the "too weak" region of the JAR scale.

[b] Values in bold are significantly different from 0 at an error level α=0.05 (XL-Stat).

Table 6

Significant model parameters of Multiple Linear Regression. (Insignificant variables are not shown. Attributes are sorted based on their *t*-values.

Source	Value	Standard error	*t*	Pr > |t|	Lower bound (95 %)	Upper bound (95 %)
Intercept	7.837	0.236	33.201	< 0.0001	7.369	8.305
FruitFlav-	1.554	0.228	6.823	< 0.0001	1.103	2.006
Aftertaste+	-1.041	0.237	-4.393	< 0.0001	-1.511	-0.571
Carbonation+	-0.751	0.274	-2.738	0.007	-1.294	-0.207
FruitFlav+	-0.926	0.347	-2.67	0.009	-1.614	-0.239
Aftertaste-	0.614	0.253	2.426	0.017	0.112	1.116
Carbonation-	0.571	0.263	2.174	0.032	0.051	1.092
Sourtaste+	-0.885	0.436	-2.029	0.045	-1.75	-0.021

Root mean square error: 1.360 on 109 degrees of freedom, Multiple R-Squared: 0.508,

F-statistic: 16.073 on 7 and 109 degrees of freedom, the overall *p*-value is < 0.0001

Table 7

Results of the PLS-R model. Attributes are sorted based on their Variable Importance (VIP) value. The upper bold line represents a VIP value of 1, while the lower bold line represents a VIP value of 0.8.

Feature	VIP	Standard deviation	Lower bound (95 %)	Upper bound (95 %)	Standardized coefficients (Feature liking):
FruitFlav-	2.133	0.251	1.641	2.626	0.280
Aftertaste+	1.616	0.274	1.079	2.152	-0.212
Sweettaste-	1.359	0.191	0.984	1.734	0.178
Odor-	1.249	0.491	0.286	2.212	0.164
Bittertaste+	1.143	0.260	0.633	1.652	-0.150
Sourtaste-	0.974	0.407	0.177	1.772	0.128
Carbonation+	0.952	0.264	0.434	1.470	-0.125
Carbonation-	0.848	0.273	0.313	1.383	0.111
Sourtaste+	0.769	0.393	-0.001	1.538	-0.101
Sweettaste+	0.682	0.480	-0.259	1.622	-0.089
Aftertaste-	0.454	0.510	-0.545	1.454	0.060
FruitFlav+	0.390	0.589	-0.764	1.545	-0.051
Color+	0.282	0.262	-0.232	0.796	0.037
Odor+	0.187	0.474	-0.742	1.116	0.025
Bittertaste-	0.127	0.270	-0.402	0.655	-0.150
Color-	0.076	0.395	-0.698	0.850	0.010

The PLS-R model used one component.

Root mean square error: 1.347 on 115 degrees of freedom,

Multiple R-Squared: 0.482, $R_{val} = \sqrt{Q^2} = 0.610$,

Intercept=7.673, Observed mean=5.812, Predicted mean=5.812.

Table 8

Results of the Penalty Analysis. The last two columns of the table contain the results of a normality test to identify which features follow normal distributions.

Feature	Level	%	Mean (liking)	Mean drops	*t*-test *p*-value[a]	Shapiro-Wilk's *p*-value[b]
Color	Too weak	30.77 %	5.806	-0.032	0.933	0.00295
	JAR	64.10 %	5.773			0.00533
	Too strong	5.13 %	6.333	-0.56		**0.55438**
Odor	Too weak	32.48 %	5.263	0.787	**0.033**	**0.26424**
	JAR	51.28 %	6.05			0.01004
	Too strong	16.24 %	6.158	-0.108		**0.55008**
Fruit flavour	Too weak	51.28 %	5.033	2.146	**< 0.0001**	**0.11198**
	JAR	33.33 %	7.179			0.00549
	Too strong	15.38 %	5.444	1.735		**0.09078**
Carbonation	Too weak	21.37 %	5.24	1.039	**0.025**	**0.1925**
	JAR	58.12 %	6.279			0.00094
	Too strong	20.51 %	5.083	1.196	**0.006**	0.00175
Sweet taste	Too weak	31.62 %	4.973	1.657	**< 0.0001**	**0.07956**
	JAR	46.15 %	6.63			0.00078
	Too strong	22.22 %	5.308	1.322	**0.001**	**0.16676**
Sour taste	Too weak	39.32 %	5.261	1.18	**0.001**	0.04608
	JAR	50.43 %	6.441			0.0042
	Too strong	10.26 %	4.833	1.607		0.02516
Bitter taste	Too weak	24.79 %	5.655	0.58	0.127	0.00755
	JAR	58.12 %	6.235			0.00109
	Too strong	17.09 %	4.6	1.635		**0.1307**
Aftertaste	Too weak	25.64 %	5.6	0.977	**0.009**	**0.44519**
	JAR	44.44 %	6.577			0.0069
	Too strong	29.91 %	4.857	1.72	< 0.0001	0.00553

[a] Values in bold are significantly different from 0 at an error level α=0.05 (XL-Stat).

[b] Values in bold indicate the non-violation of normality at an error level α=0.05 (XL-Stat).

Table 9

Comparison of all methods.

	GPCM (simple ordering, CondExact)	GPCM (significance ordering, CondExact)	GPCM (difference ordering, CondExact)	Multiple Linear Regression	Partial Least Squares Regression	Penalty Analysis
1	FruitFlav-	FruitFlav-	FruitFlav-	FruitFlav-	FruitFlav-	FruitFlav-
2	Aftertaste+	Aftertaste+	Aftertaste+	Aftertaste+	Aftertaste+	Aftertaste+
3	Bittertaste+	Bittertaste+	Bittertaste+	Carbonation-	Sweettaste-	Sweettaste-
4	Sweettaste-	Sweettaste-	Sweettaste-	Sourtaste+	Odor-	Sweettaste+
5	Sourtaste+	Sourtaste+	Sourtaste+	FruitFlav+	Bittertaste+	Carbonation+
6	Carbonation+			Carbonation+	Sourtaste-	Sourtaste-
7	Odor-			Aftertaste-	Carbonation+	Carbonation-
8	Sourtaste-			-	Carbonation-	Aftertaste-
9	Sweettaste+			-	-	-
10	Color+			-	-	-
11	Carbonation-			-	-	-

Caption to figures

Figure 1

A general five category JAR scale ranging from much too weak to much too strong.

Figure 2

Effects of Fruit flavor, Aftertaste and Sweet taste on overall liking. The “rooftop” shape of the bipolar JAR scales requires to create two one-directional variable.

Figure 3

Graphical representation of the four possible events as the basis of pair correlation method. Frequencies of cases *C* and *B* (indicated as blue on the plot) are the basis of the method.

Figure 4

Line plot summarizing the results of the three ordering methods and the three statistical tests. Crossing lines represent differences between methods and/or tests.

Figure 5

Bubble plot of the results of simple ordering and Conditional exact Fisher’s test. The vertical line divides the significant and non-significant variables. The horizontal line represents 20 % of the consumers. The size of bubbles defines the impact on liking.

Figure 6

Attributes sorted based on their variable importance. The two vertical interrupted lines indicate the Variable Importance for the Projection (VIP) value of 1 and 0.8.

Figure 7

Mean drop plot of Penalty Analysis. Mean drops are plotted against the percentage of the consumers. The vertical line represents 20 % of consumers.

Figure 1

□ Much too weak	□ Somewhat too weak	□ Just About Right	□ Somewhat too strong	□ Much too strong

Figure 2

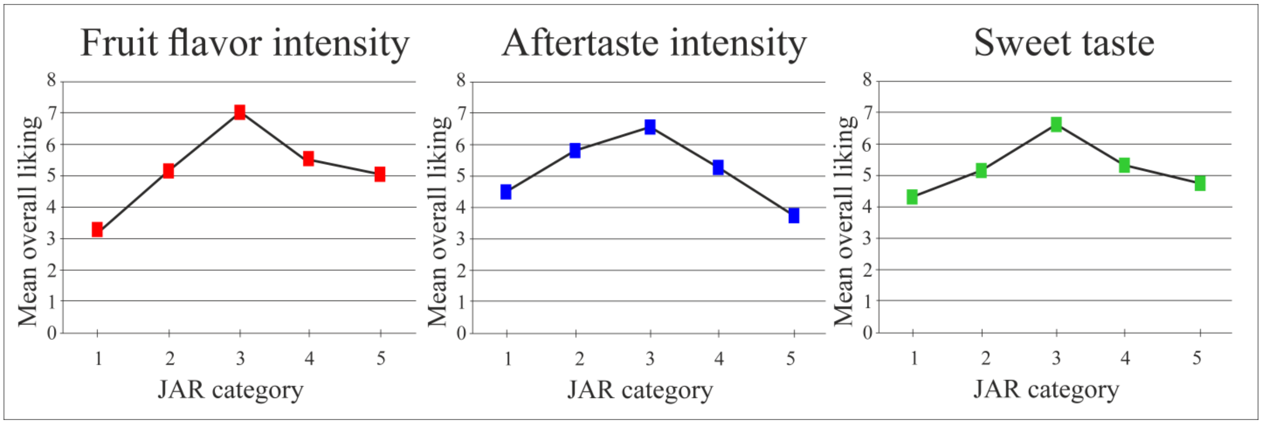


Figure 3

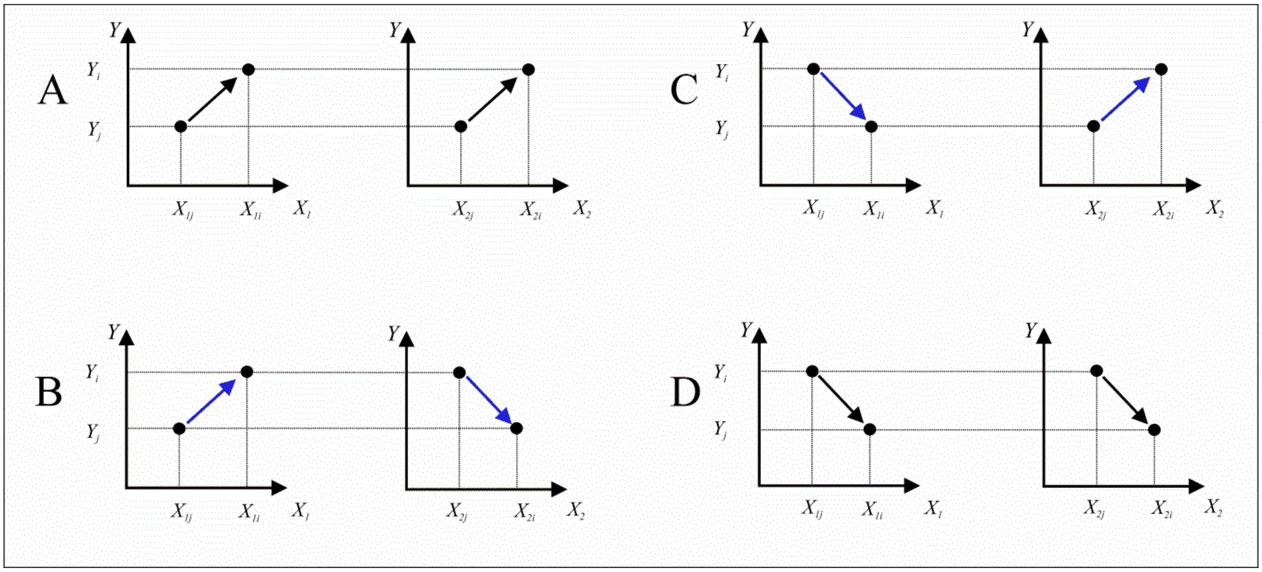


Figure 4

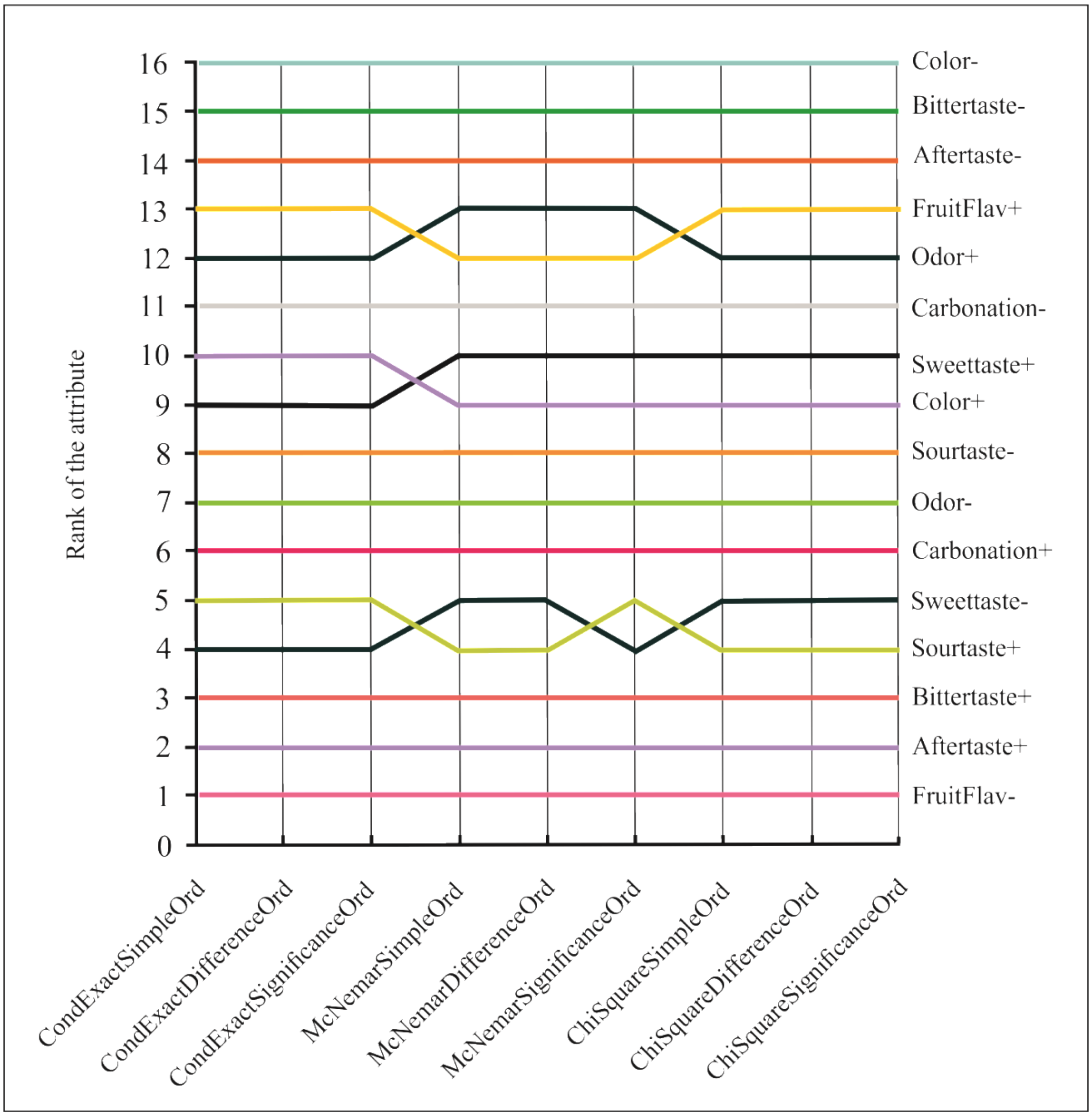
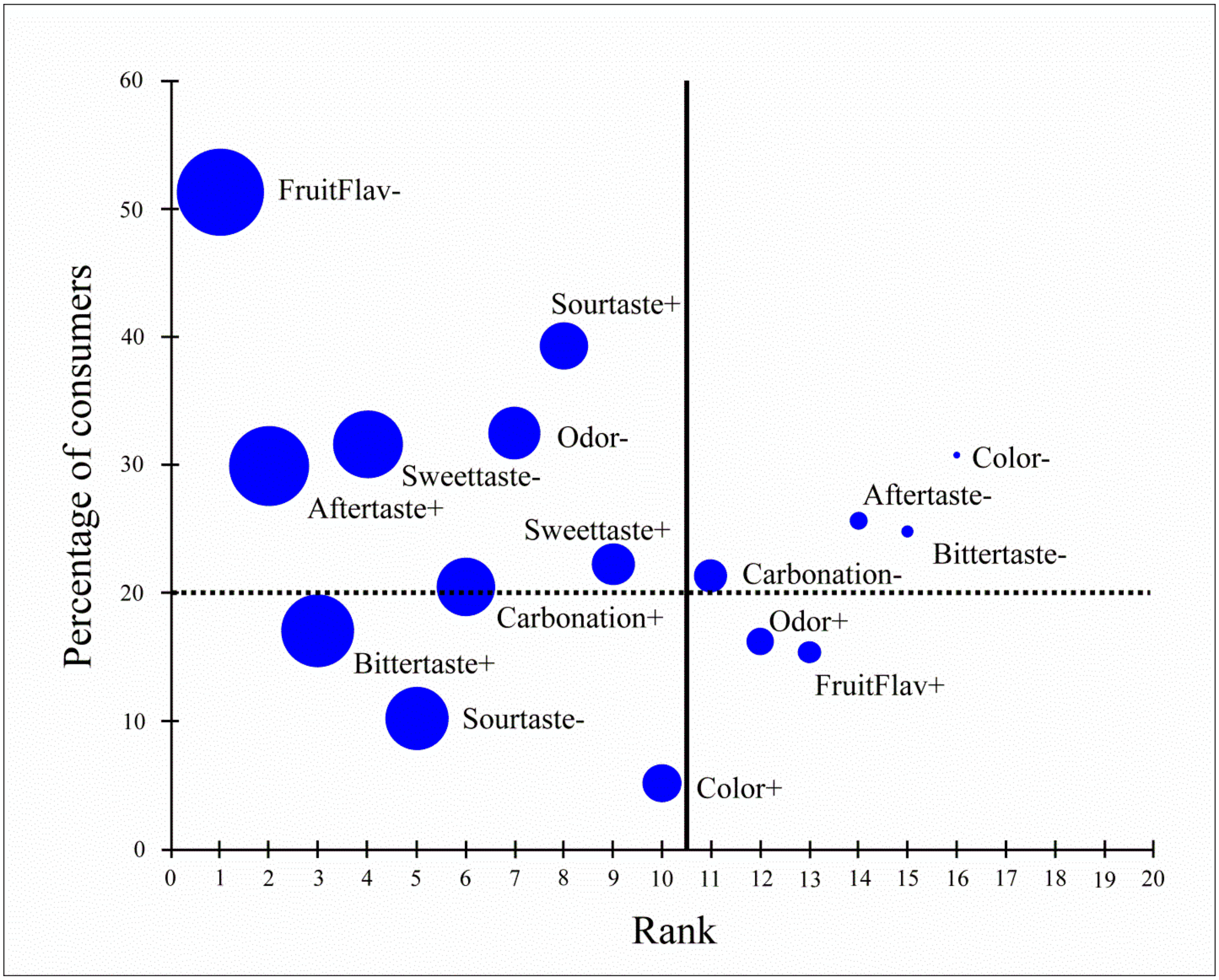

Rank of the attribute
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
Color-
Bittertaste-
Aftertaste-
FruitFlav+
Odor+
Carbonation-
Sweettaste+
Color+
Sourtaste-
Odor-
Carbonation+
Sweettaste-
Sourtaste+
Bittertaste+
Aftertaste+
FruitFlav-
CondExactSimpleOrd
CondExactDifferenceOrd
CondExactSignificanceOrd
McNemarSimpleOrd
McNemarDifferenceOrd
McNemarSignificanceOrd
ChiSquareSimpleOrd
ChiSquareDifferenceOrd
ChiSquareSignificanceOrd

Figure 5



Percentage of consumers
60
50
40
30
20
10
0
FruitFlav-
Sourtaste+
Odor-
Sweettaste-
Aftertaste+
Sweettaste+
Carbonation+
Bittertaste+
Sourtaste-
Color+
Carbonation-
Odor+
FruitFlav+
Aftertaste-
Bittertaste-
Color-
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Rank

Figure 6

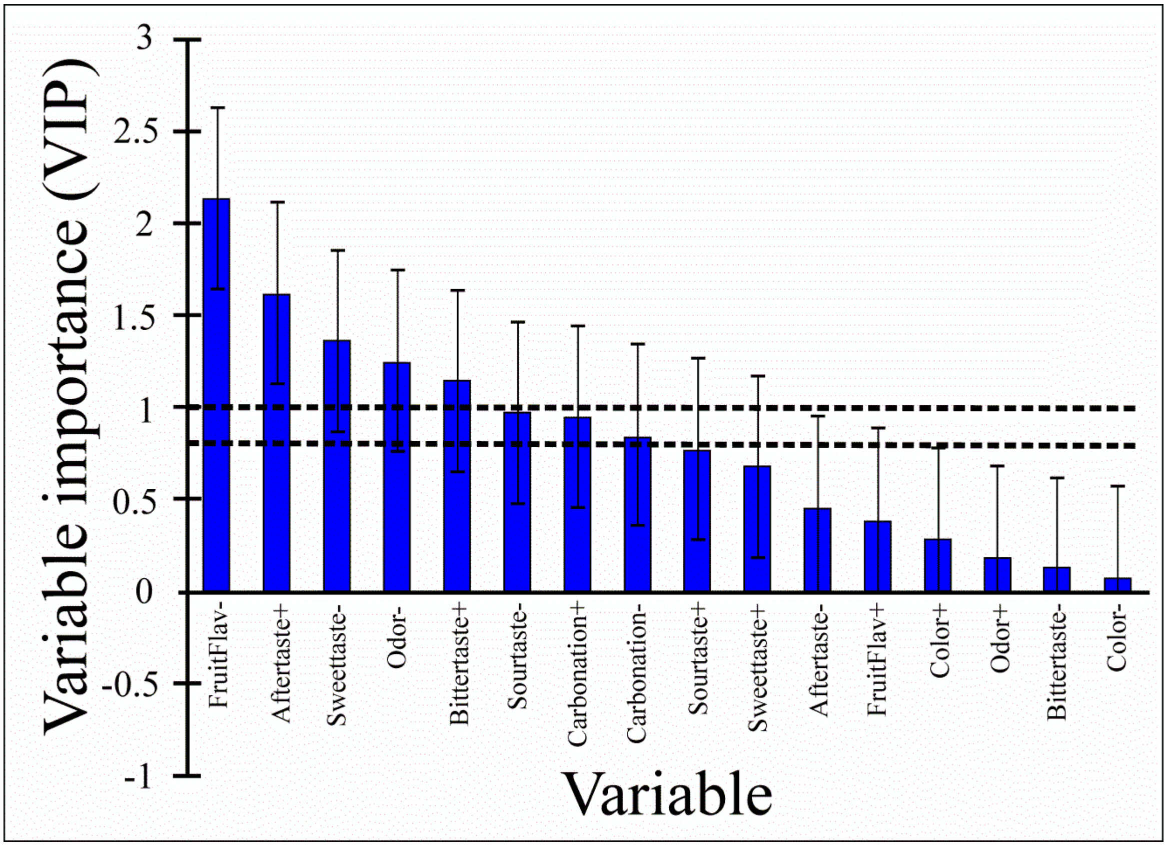
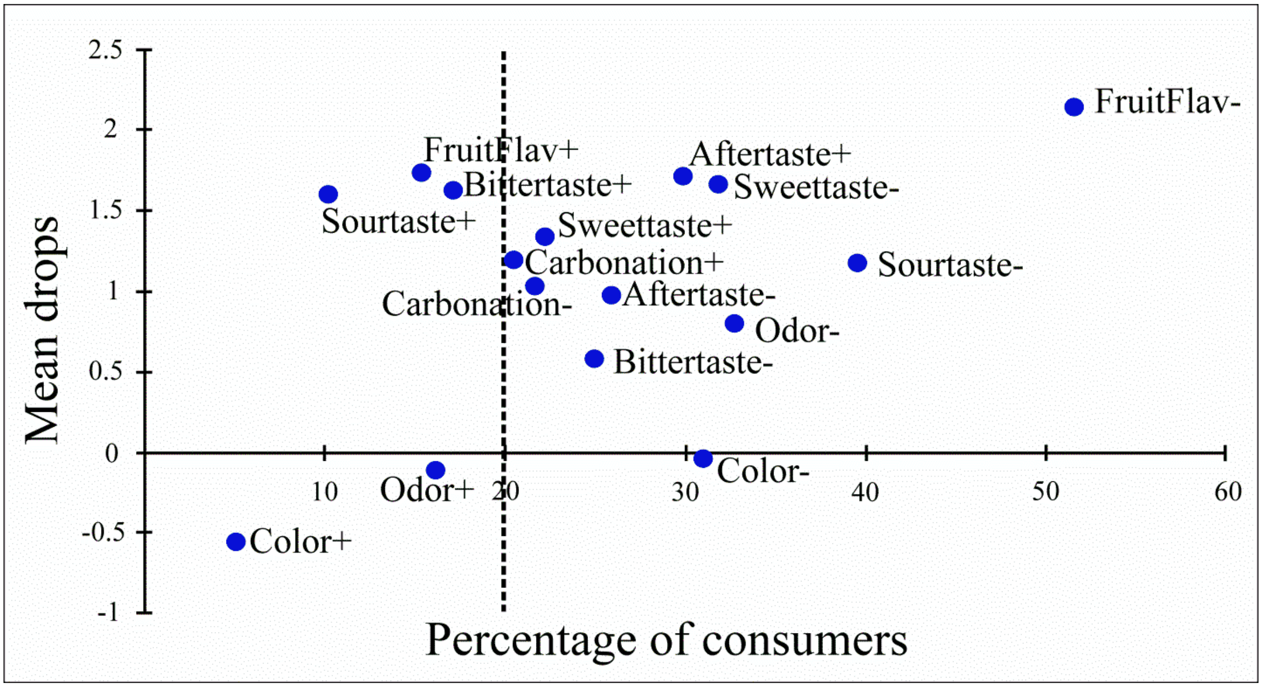

Variable importance (VIP)
3
2.5
2
1.5
1
0.5
0
-0.5
-1
FruitFlav-
Aftertaste+
Sweettaste-
Odor-
Bittertaste+
Sourtaste-
Carbonation+
Carbonation-
Sourtaste+
Sweettaste+
Aftertaste-
FruitFlav+
Color+
Odor+
Bittertaste-
Color-
Variable

Figure 7


FruitFlav-
FruitFlav+
Aftertaste+
Bittertaste+
Sweettaste-
Sourtaste+
Sweettaste+
Carbonation+
Sourtaste-
Carbonation-
Aftertaste-
Odor-
Bittertaste-
Color-
Odor+
Color+
Mean drops
Percentage of consumers
2.5
2
1.5
1
0.5
0
-0.5
-1
10
20
30
40
50
60