# COMPARISON OF THE PERFORMANCE OF A TRAINED AND AN UNTRAINED SENSORY PANEL ON SWEETCORN VARIETIES WITH THE PANELCHECK SOFTWARE

**Viktor Losó[1], Attila Gere[1], Annamária Györey[1], Zoltán Kókai[1], László Sipos[1]**

*[1]Corvinus University of Budapest, Faculty of Food Science, Postharvest Department, Sensory Laboratory, H-1118, Budapest, Villányi út, 29–43.*

***Abstract:*** In this paper the results of trained and untrained sensory panels are compared on five Hungarian commercial sweet corn samples. The two evaluations were carried out in a sensory laboratory (ISO 6658:2005), with the same experimental design, with two replicates, and the panels consisted of 10 panelists. In both cases the panels assembled the profiles of the samples according to the vocabulary chosen by the trained panelists. The results show that the untrained panel has higher standard deviation, weaker repeatability and less significant parameters (ISO/DIS 11132). However 10 of the 17 sensory attributes were significant in the case of the untrained panel, the trained panel has 15 significant parameters with lower standard deviation and good repeatability. During the statistical investigation we focused on the panel performance and used the PanelCheck open source software package to achieve this goal. We followed the workflow suggested by the researchers of the Nofima, the developers of the PanelCheck. According to the examined parameters the trained panel has better discrimination ability (F values) for attributes 'yellow color', 'hue', roughness', 'freshness', 'juiciness', 'tenderness'. There was not an attribute evaluated by the untrained panel where all the panel members reached the line representing the 5% significance level. Furthermore the trained panel has better agreement between its assessors (Tucker-1 plots) and the repeatability is much better according to the MSE plots. This examination confirms that it is necessary to train the panels in order to get reliable and consistent results.

***Keywords:*** Panel performance, PanelCheck, Sweet corn evaluation, Trained and untrained panel monitoring

## 1. Introduction

The computer aided sensory evaluation has been a topical issue of the past few years in the international industrial and scientific life. Sensory tests conducted by experts or consumers need high level IT support. On one hand the IT support is necessary during the consumer tests because of the high number of the consumers and on the other hand the expert tests need this kind of support to ensure the reliability of the sensory panel. There has been an ever growing need for the monitoring of the sensory panels. The panel leader has to know and follow the performance of the individual assessors and the panel as a whole: individual panelist's discriminating ability, individual panelist's reproducibility, individual panelist's agreement with the panel as a whole, panel discriminating ability, panel reproducibility (*Lawless – Heymann,* 2010; ISO/DIS 11132). Because of these facts have the researchers and scientists implemented two or more way mathematical methods (*Meullenet et al.* 2007, *Tomic et al.* 2007, *Martin and Lengard* 2005, *Pineau et al.* 2007, Kollár-Hunek et al. 2008, Héberger – Kollár-Hunek 2010). These methods, created to ensure the quality control and control of the panels, have been the part of different software (SAS, SensomineR, PanelCheck). The major advantage of this software is – with proper data pre-processing – that the results of the statistical methods are graphically based. In this way the performance of the sensory panels can be evaluated or analyzed rapidly and comprehensively. With this software the uni- and multivariate methods can be well combined. The panel leader gets useful information about the panel. With these important information the panelists can be trained according to their strengths and weaknesses. (*Tormod et al.* 2010, *Tomic et al.,* 2010).

## 2. Materials and methods

### *2.1 Materials*

In our study five Hungarian commercial sweet corn samples were evaluated by two panels consisting of ten panelists. The members of the untrained panel were students of the Corvinus University of Budapest, Faculty of Food Sciences. The members of the trained panel were the assessors of the Corvinus University of Budapest, Sensory Evaluation Laboratory. These trained individuals got the

training which meets the requirements of the ISO 8586-1 standard (ISO 8586-1). The preparation of the samples was conducted by the same parameters (cooking time, sample quantity, material and brand of the cooking vessels, size and temperature of the hot plate, etc). The recommendations of Beeren (2010) were followed during the sample presentation, so the small amount of samples were prepared by one person to achieve better homogeneity, and reference samples were used to reduce the standard deviation. The samples were labeled, according to the international practice (ISO 6658:2005), by 3-digit random numbers. In the literature numerous types of palate cleansers are recommended depending on the characteristic of the tested products. During our earlier experiments tap water was used, but now the much more constant composition mineral water was chosen which has neutral taste and thus does not modify the sensory properties (Aquarius) (Sipos, 2010).

The sessions were conducted in two different days one for the trained and one for the untrained panels, with two replicates per sessions. The sensory evaluations were conducted in a laboratory which meets the ISO 8589:2007 requirements (ISO 8589:2007). The 17 measured attribute were the following: Yellow color, Hue, Size, Roughness, Freshness, Odor intensity, Cooked corn odor, Sweet odor, Texture, Juiciness, Skin chewiness, Tenderness, Global taste intensity, Sweet taste, Salty taste, Cooked taste and Aftertaste.

### *2.2 Methods*

In our research the performance of a trained and an untrained panel were evaluated according to the workflow offered by the creators of the PanelCheck. The results of the untrained and the trained panels were compared by uni- and multivariate mathematical-statistical methods. The methods will be presented in an order that complies with the suggested data analysis workflow. The sensory test was carried out by the means of profile analysis (ISO 13299:2003). Results of the tests were summarized through spider plots. These figures show the similarities and differences of the tested samples across the attributes.

## Multivariate statistical methods

#### 2.2.1 Mixed modell ANOVA

As a first step, mixed model ANOVA was conducted for assessing the importance of the applied sensory attributes in detecting significant differences between the samples. The main reason for using this method is to eliminate the unimportant attributes from the further analysis. The method is based on modeling samples, assessors and their interactions in two-way ANOVA model or samples, assessors, replicates and their interactions in three-way ANOVA model, and then testing for the sample effect by regular F tests (on the vertical axis are represented the F values). In each case, the assessor and interaction effects are considered random. Only attributes that are significant at the chosen significance level (in our case we defined 5%) for product effect are considered as the subjects of further analysis.

#### 2.2.2 Tucker

In the next step, the multivariate analysis method Tucker-1 was applied in order to get an overview over assessor and panel performance using multiple attributes. The essence of Tucker-1 method is that a PCA is applied on an unfolded data matrix. This data matrix consists all individual matrices (Xi,) ordered horizontally. This unfolded matrix then consists of J rows, where each row represents the average across replicates, and IXK columns, where I represents the number of assessors and K represents the number of attributes. This means that the dimension of the unfolded matrix will be J X (I*K). In the case of our data set the dimension would be 5 X (10*17), with J = 5 samples, and K = 17 attributes and I = 10 assessors in both cases.

As a result the method will provide two different types of plots: a common scores plot and a correlation loadings plot. The common scores plot shows how the tested J samples relate to each other, i.e. it visualizes similarities and dissimilarities between the samples along the found PC's. This plot gives no direct information on assessor or panel performance, but it can be used as a useful tool to investigate the discrimination ability of the panel taking the explained variances into account. Based on the high explained variance in the first few PC's (it means that there is a systematic variation in the data set) the discrimination ability of the panel can be considered as good.

The correlation loadings plot provides performance information on individual assessors and the sensory panel as a whole. The plot contains I*K dots, with each dot representing one assessor-attribute combination. The specific dots of one assessor or one attribute are highlighted in order to visualize the performance. The information about the performance comes from the position of the dots. If the attribute or assessor contains a lot of noise the dots will located close to the origo. On the other hand the dots will position around the 100% variance ellipse if there are more systematic variation. The other ellipse represents the 50% explained variance. If the dots of an assessor of an attribute fall under this line the performance will considered as not enough good. For a well-trained and calibrated panel the correlation loadings of the attribute under investigation should be close to the outer ellipse with all panelists clustered closely together.

#### 2.2.3 Manhattan diagram

Manhattan plots in general provide an alternative way to visualize systematic variation in data sets. These plots are easy to look at and provide useful information for screening

purposes. The information visualized by Manhattan plots may be computed with different statistical methods. In this paper, we used PanelCheck to provide these kind of plots. In our study I*K explained variances will be given. The number of the assessors in both panels is I = 10, the number of explained variances is 10*17 with the 17 attributes.

Manhattan plots are easy to understand because it uses the shades of grey to visualize the explained variances and the PC's. The principal components located on the vertical axis. The shades have to endpoints, one is the black which represents the 0% explained variance, and the other is the white which means the 100% explained variance. The lighter the plot of an assessor or attribute, the better the performance is. Typically, the color will be darker for PC1 and then get lighter with each additional PC. In other words, the explained variance at PC3 is the sum of the explained variances of PC1, PC2 and PC3 (Tomic, O et al. 2010). In our case the method is most useful when the explained variances for the different attributes are presented in separate plots. The other option is to sort the explained variances by assessors in order to obtain information about the differences between panelists. The shades help the user to identify these differences and after this one can choose the specific methods to investigate deeper the chosen assessor or attribute. This can shorten the time of analyzing performance. Both of the methods mentioned above are implemented in PanelCheck but we have chosen the plots sorted by attributes because we were interested in the differences between panels according to the measured attributes. For further analysis the other option can be used later in another study to improve the trained panel.

## Univariate statistical methods

### 2.2.4 F plots

F values can be used to check each assessor's ability to detect differences between the samples for a given attribute. F values can be plotted in a bar diagram, giving an overview over the performances of all assessors in the panel. Generally speaking, it can be said that the higher the F value of an individual assessor for a certain attribute, the greater is the ability of this assessor to distinguish between the measured samples. Besides the F values there are two horizontal lines in the F plots. These lines represent the 1% and the 5% significance level. If there are differences between the tested samples one can expect the assessors to obtain higher F values than the level of the 1% and 5% significance level.

### 2.2.5 MSE plots

The MSE values are the mean square errors (random error variance estimates) from the one-way ANOVA model so they provide a direct measure for the repeatability of the individual assessors. Similar to the F values a total of I*K MSE values are calculated and plotted in a bar chart. If an assessor almost perfectly repeats the results, this MSE value should be close to zero. This means that in contrast to the F values the lower the MSE values mean the perfect performance. Generally speaking, the lower the MSE value, the better the repeatability of the particular assessor. It should be used with the F values to get a realistic overview over the panel's performance. If differences between the samples are given, an assessor should ideally have high F values and low MSE values.

### 2.2.6 p*MSE plots

In a p*MSE plot the assessor's ability to detect differences between samples (p) is plotted against their repeatability (MSE). A total of I*K pairs of p and MSE values are computed and plotted in a scatter plot. They can be presented together in various ways (for instance all at the same time, only for one attribute at a time or only for one assessor at a time) and with highlighting of the assessors or attributes that one is particularly interested in. The perfect p*MSE plot has low p and MSE values close to the zero and all of the dots are clustered around the origo. However it is true in the case the difference is really present between the tested samples. The p*MSE plot is a valuable tool to quickly and easily detect which assessors perform poorly for a certain attribute. A great advantage of the p*MSE plot is that it displays distinguishing performance and repeatability in a single plot for all assessors and all attributes. That means that with a single plot one can get a quick overview over the performance of the entire panel.

### *2.3 Preparation of data*

There is always a data preparation step before importing data into PanelCheck. The software requires a data structure, so one needs to order the data sets according to this structure. The most obvious way of doing this is the use of the Microsoft Excel, where the assessors should be ordered in the left column (A), and the attributes should be on the first row of the table. During the data import the software asks which columns are the columns of the „assessors", „replicates" and „samples". Furthermore it offers an opportunity to choose attributes or assessors in which one is not interested in so one can leave out these objects. After this the data import is done. During the process the software automatically checks the data sets and gives a „Summary" which contains the parameters of the data set. With this tool one can see that all of the data are OK.

## 3 Results and discussion

The plots used during the analysis have unique information and can be used independent on the other graphical methods. But it is much more effective to use them as a collection of methods, so they can complete each other. Following these the panel leader can have a more

comprehensive picture over the data set. With this information in hand the weaknesses of strengths of the panel can be analyzed more efficient. The methods and plots used in this paper are based on two replicates. According to the results of the trained panel we used the workflow indicated as gray in Fig. 1 and the results are discussed detailed below.

### *3.1 Trained and untrained panel, 2-way ANOVA*

According to the workflow in Fig.1 a 2-way ANOVA was applied on the data sets of the two panels as the first step of the analysis. The results are visualized on Fig. 2 and 3. The plots have three different colors. The red means the significance level is $p<0.001$, the orange means $p<0.01$, the yellow means $p<0.05$ and the gray means the non-significance. The non-significant attributes were excluded from the further analysis. According to this the trained panel has two non-significant parameters, Roughness and Salty taste, and the untrained has four, Size, Sweet odor, Texture and Aftertaste. So these attributes had been excluded from the further analysis. The untrained panel has twice as many attributes which are non-significant. Although the results of trained panel are not fully satisfying but they can be improved by further training and practice. It is important to mention that the trained panel had no specific product training because they have to work with different kind of products.

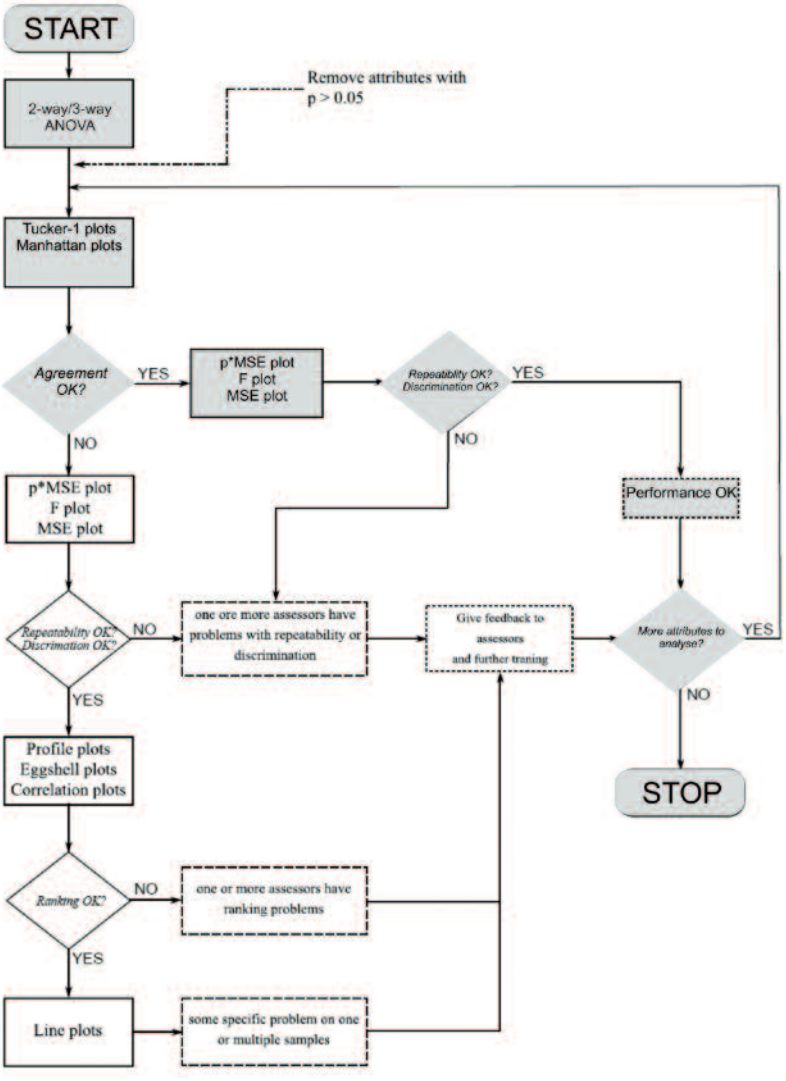


***Fig. 1:*** The applied workflow indicated as gray

### *3.2 Trained and untrained panel, Tucker-1 plots*

The next step is the applying of multivariate statistical methods to get an overview about the performance of the two panels. The panel has good performance if the members of the panel (dots representing them) are close to the outer ellipse (100% explained variance), and to each other (Tormod et al. 2010). Fig. 4 represents the Tucker-1 plots of the untrained panel. It can be seen that the dots are scattered at most of the attributes so the panel has weak performance. There is only one attribute (Sweet taste) where all of the panel members are between the two ellipses, but the assessors are far away from each other so the panel agreement is weak. At the juiciness attribute the assessors are close together but one of them is located under the 50% ellipse so his or her variance is low. If his or her results would be excluded this performance could be considered as good. The results of the trained panel on the Fig 5 show better panel agreement. The assessors have lower variance at the Cooked odor, Odor intensity and Aftertaste attributes. Furthermore there are some assessors who have lower variance at the Size, Sweet odor, Texture and Cooked taste attributes. The lack of agreement is not only because of the panel's weakness, it is possible that there were no differences between the tested samples, or if there was it was only a little difference. As a result it can be said that the trained panel has better agreement; there were more attributes with excellent results. In contrast the untrained panel has no parameter with excellent results.

### *3.3 Manhattan plots of the trained and untrained panels*

Using the Manhattan plots the systematic variance of one specific attribute can be analyzed. The performance of the two panels can be compared according to this. In this case we selected the visualization by attributes because we are interested in the differences of the panels. In the case of seeking the performance differences between the assessors of one specific panel one should select the visualization by assessors. Both panel needed

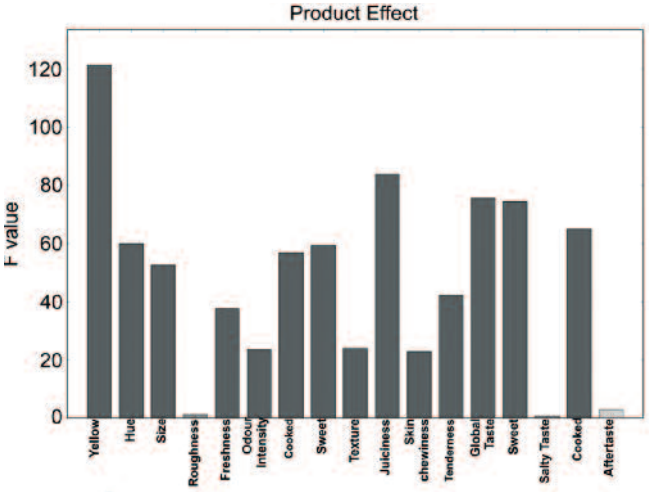


***Fig. 2:*** ANOVA of the trained panel. Two parameters do not reach the p<0.005 significance level (Roughness and Salty taste)

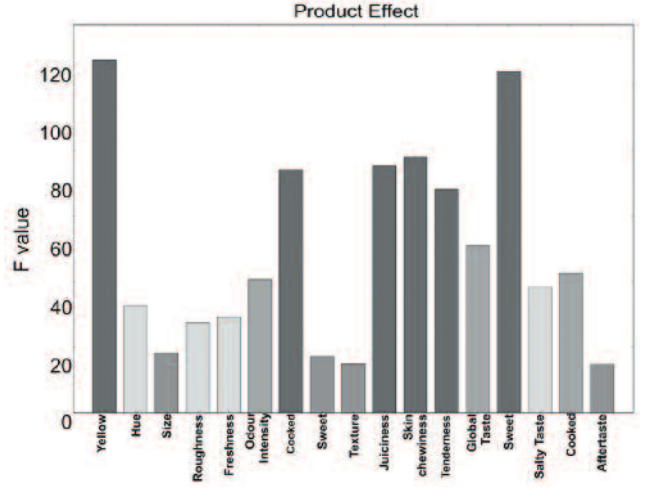


***Fig. 3:*** ANOVA of the untrained panel. Only six parameters reach the p<0.005 significance level and there are four parameters with p<0.001 level (Size, Sweet odor, Texture and Aftertaste)

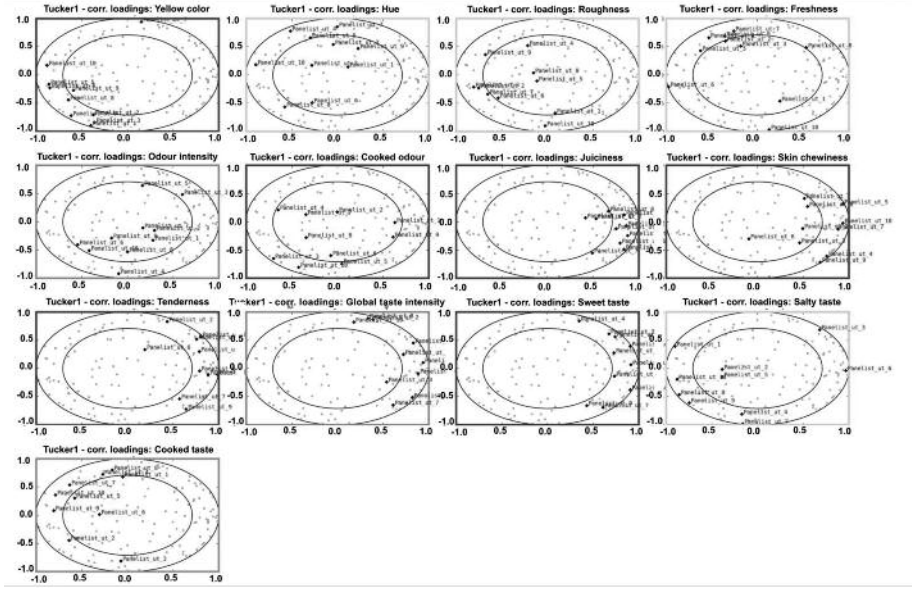


***Fig. 4:*** Tucker-1 plot of the untrained panel

more PC's to reach higher explained variance. The best results of the trained panel were at the Skin chewiness and Tenderness. According to the results the untrained panel has lower explained variance at the attributes as the trained panel. It can be stated that trained assessors used and understood better the attributes as untrained panel (4 out of 17 were nearly the same, and 11 out of the 17 were better understood and used).

### *3.4 F plots of the trained and untrained panels*

When analyzing the F plots it has to be considered that the higher F values mean the better discrimination ability. Furthermore the significance levels are plotted too. More trained as untrained assessor reached the 5% line, as it can be seen in Fig. 6 and 7. Besides this the F values of the trained panelists were higher than that of the untrained panel. Nevertheless there are trained assessors who had weaker performance at specific attributes. With the PanelCheck one can find the reasons of this weaker performance and can suggest further trainings which opportunity could be a good base of a further study.

There are no untrained assessors who could reach the 1% significance level at all of the measured attributes. In contrast there were only four trained assessor who could not reach this 1% significance line at one of all attributes. According to the results of the F plots the trained panel has better discrimination ability than the untrained panel.

### *3.5 MSE plots of the trained and untrained panel*

After analyzing the discrimination ability the next step is to measure the repeatability during our way to explore the performance of the two panels. For this purpose the MSE plots are the best tools. The results of the MSE plots can be summarized in one sentence: the lower the MSE value the better the repeatability. The better the repeatability of the assessor the closer the MSE value will be to zero.

Two panelists out of the untrained panel, panelist 2 and 4, have as low MSE values as the average of the trained panel. They probably have good sensory abilities, or good sensory memory, since they gave results similar to the trained panel with the lack of training. But their F values were weaker than that of the trained panel.

The MSE values of the trained panel were lower than that of the untrained panel but the untrained panel has some better results. It would be efficient to analyze some attributes to

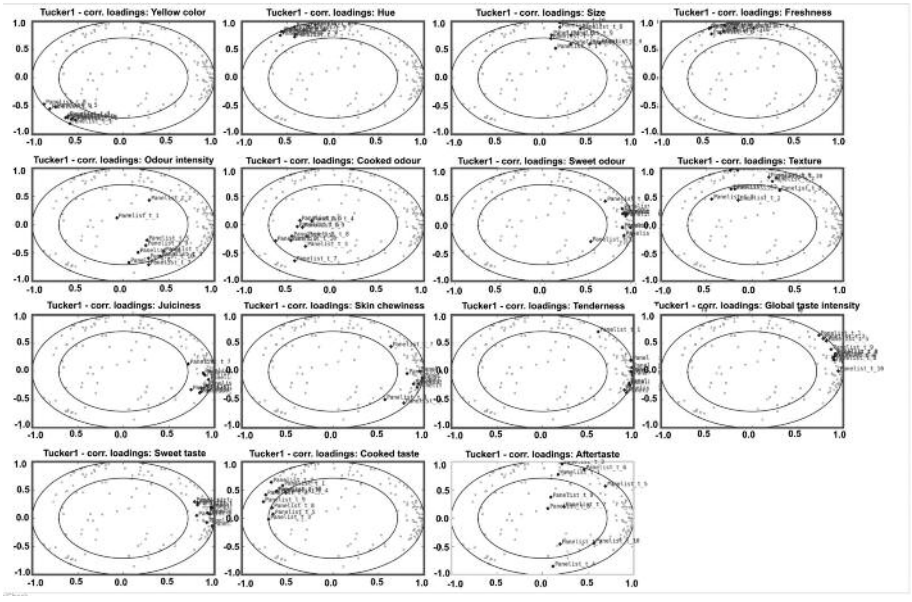


*Fig. 5:* Tucker-1 plot of the trained panel

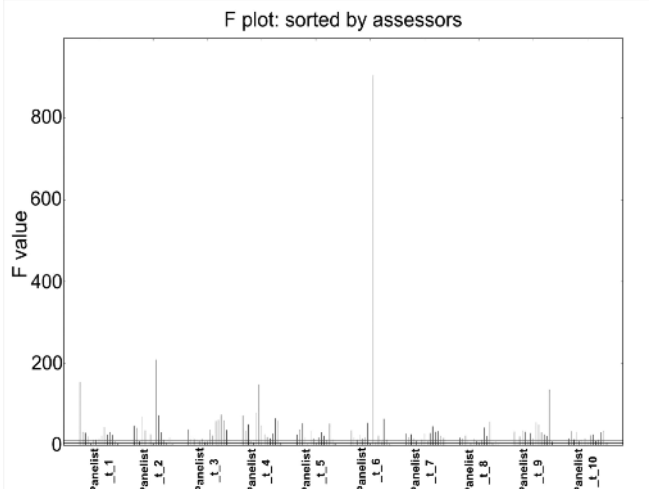


*Fig. 6:* F plot of the trained panel

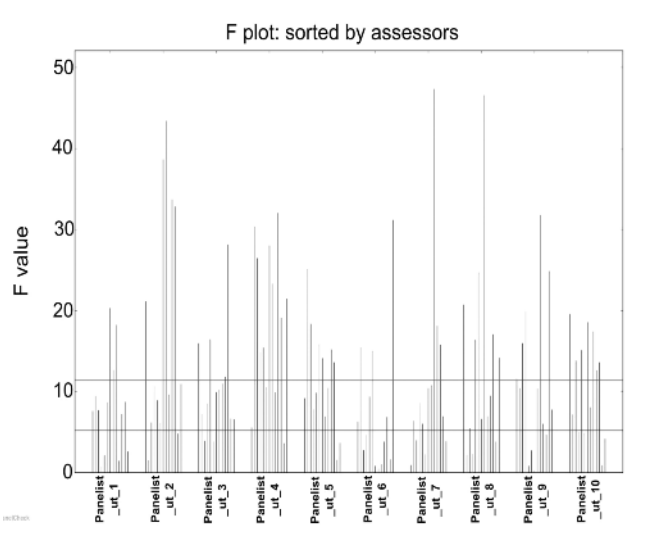


*Fig. 7:* F plot of the untrained panel

achieve better results, or with other words: to achieve better repeatability.

### *3.6 p*MSE plots of the trained and untrained panels*

The p values of the trained panel were lower (the highest of them was p=0.4, in contrast the highest of the untrained panel's p values was p=0.8). The same can be said about the MSE values so the trained assessors have lower results (the MSE values of the trained assessors were around 20 with some around 60, but the untrained panel's results were around 50 with some around 300), which means that the trained panel has better discri-

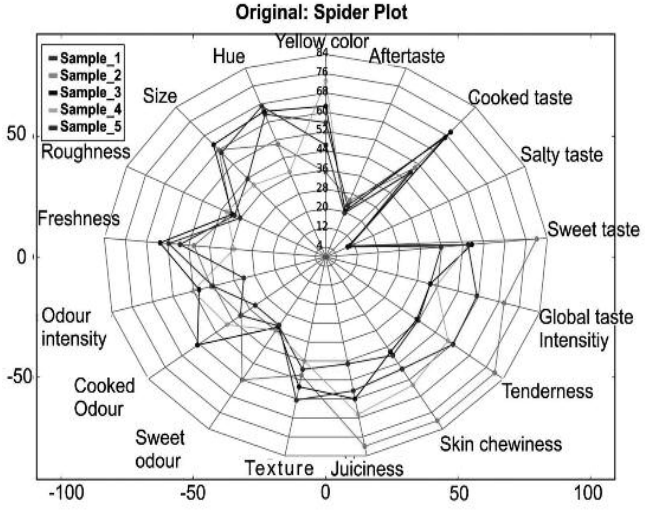


*Fig. 8:* Spider web plot of the trained panel

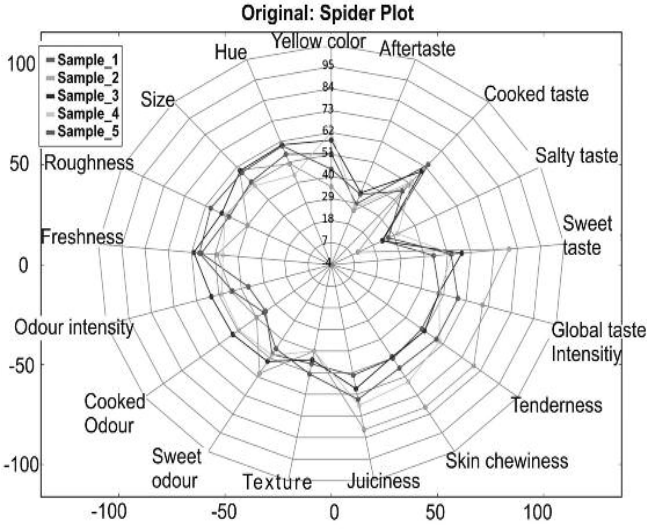


*Fig. 9:* Spider web plot of the untrained panel

mination ability. According to the results the trained panel has better panel agreement because the standard deviation of their MSE values was lower.

### *3.7 Spider plots of the panels*

In the following we would like to present the profile plots of the two panels. Fig. 8 represents the plot of the trained panel and Fig. 9 is the profile plot of the untrained panel. There are some differences between the two panels.

According to the plots the main differences were the Sweet taste, Global taste intensity, Tenderness, Juiciness and Skin chewiness. On the plot of the trained panel one can identify bigger differences so the trained panel was more sensitive.

## 4 Conclusions

After the analysis it can be said that according to all of the applied methods the results of the trained panel were better than that of the untrained panel so the training, validation and monitoring of the panels are important during sensory evaluations. For the two panels the most difficult attribute to use were Cooked odor and Odor intensity. These two attributes have the highest variations in the Tucker-1 plots and they required the most PC's in the Manhattan plots to reach higher level of explained variance. Both attributes are based on olfaction so it can be said that it would be useful to train the odor detecting ability of the panels. It could be done using reference samples, and after this training the panels could reach better results.

In general trained panel has a good performance but the F values of panelist 2 do not reach the 5% significance level at 5 out of 15 attributes which is high among the trained assessors. Furthermore the MSE values of the assessor are high among the other trained assessor's MSE values. So the assessor's discrimination ability and repeatability do not match to the average of the trained panel. It would be very useful to analyze deeper the results of panelist 2 or conduct more tests to find out what was the problem. It could be a temporarily or a permanent problem. Having this information about the panelist the panel leader can make a decision about the results of the assessor. They can be excluded or the panelist can trained to get better results.

Among the untrained panel there were panelists having very good performance (panelist 2 and 4). Their F values were not very different from the average of the untrained panel but the MSE values of panelist 2 were absolutely satisfying. Although panelist 4 could not repeat the results of Salty taste but this is the only value which is too high, so it could be the reason of a mistyping and not the weakness of his or her tasting. These two assessors could be the member of the trained panel after adequate results of the proper trainings.

The spider web plots showed that the trained panel discriminated the products more effectively.

## 5 References

**Beeren, C.J.M. (2010):** Establishing product sensory specifications. In: Kilécast, D. (ed.) (2002): Sensory analysis for food and beverage quality control. Woodhead, Cambridge. 75–96.

**Héberger, K. & Kollár-Hunek, K. (2011):** Sum of ranking differences for method discrimination and its validation: comparison of ranks with random numbers. J. Chemometrics, pp. 151–158.

**ISO 8586-1** – Sensory analysis – Methodology – General guidance for the selection, training and monitoring of assessors and experts

**ISO 8589:2007:** Sensory analysis – General guidance for the design of test rooms.

**ISO/DIS 11132** Sensory analysis – Methodology – General guidance for monitoring the performance of a quantitative sensory panel.

**ISO 13299:2003** Sensory analysis – Methodology – General guidance for establishing a sensory profile

**Kollár-Hunek, K., Heszberger, J., Kókai, Z., Láng-Lázi M. & Papp, E. (2008):** Testing panel consistency with GCAP method in food profile analysis. Journal of Chemometrics. pp. 218–226.

**Lawless, T.H. & Heymann, H. (2010):** Sensory evaluation of food. spen Publisher 2nd edition, Gaithersburg, Maryland. pp. 244-247.

**Martin, K. & Lengard, V. (2005):** Assessing the performance of a sensory panel: Panelist monitoring and tracking. Journal of Chemometrics, 19, 154–161.

**Meullenet, J-F., Xiong, R. & Findlay, C. F. (2007):** Multivariate and Probabilistic Analyses of Sensory Science Problems. Wiley-Blackwell, New York, NY. PP. 27-47.

P**ineau, N., Chabanet, C. & Schlich, P. (2007):** Modeling the evolution of the performance of a sensory panel: A mixed-model and control chart approach. Journal of Sensory Studies, 22, 212–241.

**Sipos, L. (2009):** Ásványvíz-fogyasztási szokások elemzése és ásványvizek érzékszervi vizsgálata. PhD értekezés. Budapesti Corvinus Egyetem, Döntéstámogató Rendszerek Doktori Iskola, pp. 110.

**Tomic, O., Luciano, G., Nilsen, A., Hyldig, G., Lorensen, K., Næs, T. (2010):** Analysing sensory panel performance in proficiency tests using the PanelCheck software, European Food Research and Technology, Vol 230 pp 4

**Tomic, O., Nilsen, A. N., Martens, M., Næs, T. (2007):** LWT - Food Science and Technology, Vol 40 pp. 262–269.

Tomic, O., Nilsen, A., Martens, M. & Næs, T. 2007. Visualization of sensory profiling data for performance monitoring. LWT- Food Science and Technology, 40, 262–269.

**Tormod, N., Per, B.B. & Tomic, O. (2010):** Statistics for sensory and consumer science. Wiley, Chicester. pp. 11–34, 193–245.