Which Just-About-Right feature should be changed if evaluations deviate? A case study using sum of ranking differences

Attila Gere[a], László Sipos[a], Sándor Kovács[b], Zoltán Kókai[a], Károly Héberger[c,]*,

[a] Szent István University, Faculty of Food Science, Sensory Laboratory,
H-1118 Budapest, Villányi út 29-43 Hungary
[b] University of Debrecen, Faculty of Economics and Business, Institute of Sectorial Economics and Methodology, Department of Research Methods and Statistics, H-4032 Debrecen, Böszörményi út 138. Hungary

[c] Plasma Chemistry Research Group, Institute of Materials and Environmental Chemistry
Research Centre for Natural Sciences, Hungarian Academy of Sciences,
H-1117 Budapest, Magyar tudósok körútja 2 Hungary

CITE THIS:
Attila Gere, László Sipos, Sándor Kovács, Zoltán Kókai, Károly Héberger*,
Which JAR feature should be changed if evaluations deviate? A case study using sum of ranking differences
Chemometrics Intelligent Laboratory systems, 161 /-/ 130-135 (2017)

http://dx.doi.org/10.1016/j.chemolab.2016.12.007

Available online 13 December 2016

* To whom correspondence should be addressed.
Phone: (+36-1) 3826509; E-mail: heberger.karoly@ttk.mta.hu

Abstract

Several different approaches have been introduced for analysis of just-about-right (JAR) data; however, their results are sometimes deviating or even contradictory. More reliable results are gained, if a consensus of many methods is determined. A specific approach is presented to compare and select JAR attributes of food products. Overall liking was set as dependent (Y) variable and the JAR variables were used as independent (X) variables for regression methods. The mean drop value, difference between the mean overall liking of the attribute as optimum (or JAR) and that of mean overall liking of the attribute as an extreme, *i.e*. too much and not enough, was used for penalty analysis and its variants. Generalized Pair Correlation Method (GPCM) compares the impact of the JAR variables on overall liking pairwise and the probability weighted difference ordering was applied for ordering the attributes. A special data fusion is suggested based on the sum of ranking differences (SRD), primarily developed for method comparison. SRD method was able to rank the JAR variables based on their differences from a benchmark defined by all of the JAR evaluation methods in maximal performance. This enables also to group the product attributes. Moreover, it gives recommendations for how to optimize the products based on the results of several JAR methods and helps to gain a more reliable evaluation and selection of JAR attributes. The significant features can be identified easily when the SRD procedure is completed by the frequencies of consumer evaluations. The same data matrix transposed is suitable to rank the evaluation methods using the average of all evaluation methods (consensus). From among the JAR evaluation techniques, GPCM proved to be closest to the average, *i.e.* it can be used for substitution of the other techniques.

Keywords: Just-about-right scale; attribute selection; product optimization; ranking of JAR attributes; data fusion, method evaluation

1. Introduction

Just-About-Right (JAR) scales are frequently used in product development. The midpoint of this bipolar scale is labeled as “just about right” and the two ends are semantic opposites, for example, “not salty enough” and “too salty” [1,2]. JAR scales have an odd number of categories (usually between three and nine). In sensory science, JAR scales are often used with untrained panelists (or consumers) to unfold the strengths and weaknesses of a product. In such a case, the hedonic scores are also measured; hence, the effect of the JAR attributes on the liking can be analyzed. Several methods have been introduced in the literature for assessing JAR evaluations. Essentially, there are two different approaches: i) methods without hedonic scores (*i.e*. comparison of the JAR evaluations of products) and ii) methods taking into account the hedonic scores.

The following techniques can be enumerated in the first group i): graphical methods (graphical data are displayed and graphical scaling), calculations of percent differences from a given standard sample and/or JAR, computations of mean, mean direction or mean absolute deviation, Student’s *t*-test (one sample) and further multivariate methods (biplots, correspondence analysis, principal components analysis, *etc*.) [3], Chi-square test (when comparing JAR distributions of products), Cochran-Mantel-Haenszel test [4], Stuart-Maxwell test, McNemar’s test, Student’s *t*-test and analysis of variance [3], proportional odds/hazards models [5], Thurstonian ideal point modeling [6] and signal-to-noise ratio model [7].

The second group contains methods, which take into account the hedonic ratings, such as the widely used penalty analysis and its different modifications (penalty analysis using the mean of the proportion of respondents who scored the product JAR, *etc*.), opportunity analysis [3], PRIMO analysis [3], bootstrapping penalty analysis [5], ordinary least-squares regression (OLS) [8], Chi-square test (determining whether the consumers find the product lower than the JAR score), Spearman’s rank correlation coefficient, multiple linear regression [3], multivariate adaptive regression splines (MARS) [5], partial least squares regression using dummy variables [9] and generalized pair-wise correlation method (GPCM) [10,11].

Furthermore, the second group of methods can be differentiated based on the way of calculation of the impact of the JAR attributes on the liking scores. One part of the methods assesses the JAR variables completely, and the other part divides each JAR variable into two parts, where the first part belongs to the “too low” while the second part belongs to the “too much” region (*e.g*. penalty analysis).

Several methods have been introduced to assess the relative importance of predictors, which can successfully be used to identify drivers of consumers liking, such as Lindeman, Merenda and Gold’s method [12], Breiman’s Random Forests [13] and Johnson’s relative weight algorithm [14], *etc*. However, these methods have only been used to assess the role (weight) of different liking and not assessing the JAR attributes on overall liking [15].

The aim of this work is to elaborate an approach, which uses a bunch of multivariate statistical methods to identify the key JAR variables for product development. This way the identification of JAR variables becomes more reliable. Furthermore, it will be introduced how to evaluate the connection between JAR and hedonic data using multiple methods on the example of product “170” from the data set “ASTM MNL63”. Another aim is also formulated: to select the best evaluation method(s) for JAR analysis.

To achieve our goal, the procedure of sum of ranking differences (SRD) was applied. SRD is a quick, simple and general technique suitable to compare methods or statistical models fairly as well as to rank them based on their similarities and/or differences [16]. It is easy to use and the final result is a unique ranking (and grouping) validated by correct statistical tests. The SRD method has been applied in several fields and by numerous authors (*e.g.*, for column selection in chromatography [16], for sensory panel testing [17,18], for prevention of over-fitting in PLS calibration [19]. The SRD method was used to evaluate proficiency tests along with principal component and cluster analysis [20], recently, the equivalency of SRD with multicriteria decision making was proven [21]. Furthermore, a possible alternative of the method was introduced by Koziol [22]. Despite of its popularity in chemometrics, there has not been any attempt to identify JAR product attributes based on the optimization of different methods.

2. Materials and Methods

2.1 Materials

The data set provided by ASTM MNL-63 was used, which consisted of the evaluations of five products using six JAR variables (size, color, amount of flavor, amount of salt, thickness and stickiness) along with one overall liking variable. In the following, the results of product "170" will be introduced and discussed.

2.2 Just-About-Right (JAR) data analysis methods

Those JAR data analysis methods were chosen, which divide the original variables into two parts and take into account the hedonic scores. These methods determine not only the impact of the significant variables on liking but the direction of the further product development (*i.e.* if a given attribute is too strong, then, reduction of the intensity gives higher consumer acceptance). Table 1 lists the chosen methods provided by ASTM MNL-63 standard.

Table 1 Methods for evaluation of JAR scales and their parameters.

Name	Abbreviation	Used parameter
ordinary least-squares regression	OLS	*t*-values of individual parameters
penalty analysis	Penalty	mean drop values
bootstrapping penalty analysis	bPenalty	mean drop values
generalized pair correlation method	GPCM	number of pWinner values
partial least squares regression using dummy variables as dependent variable (Y)	PLS-dummy	*t*-values of parameters
multiple linear regression	MLR	*t*-values of parameters
penalty analysis for JAR mean method	wPAforJARMean	mean drop values
weighted penalty analysis for grand mean method	wPAforGrandMean	mean drop values

Details of the methods listed in Table 1 are described in ASTM MNL-63 and GPCM is described by Heberger and Rajko in detail [23].

The aim of JAR data analysis is to assess the impact of JAR variables on consumer liking. However, this is done differently by the applied methods. In the case of regression methods (OLS, PLS-dummy and MLR), overall liking was set as dependent (Y) variable and the JAR variables were used as independent, explanatory (X) variables. Hence, the *t*-values of individual parameter estimations were applied to assess the significance of attributes. OLS was done separately for each variable (one at a time), while PLS-dummy and MLR used all variables to model overall liking.

Penalty analysis (and its variants denoted by bPenalty, wPAforJARMean and wPAforGrandMean) use the so-called mean drop values which is calculated as the difference between the mean overall liking of the consumer group, who rated the attribute as optimum (or JAR) and the mean overall liking value of those who rated the endpoint of the attribute (*i.e.*: too much or not enough). In case of bPenalty, 1000 bootstrapped data matrices were generated and penalty analysis was run on each of them separately. Then, their average mean drop values were calculated. The notation wPAforJARMean stands for a modified penalty analysis, in which the mean drops were weighted by the number of the JAR group members. Mean value of all the respondents was subtracted from the mean liking of the JAR groups (wPAforGrandMean) instead of the mean liking of those who rated the product as JAR (as it was done in penalty analysis).

GPCM compares the impact of the JAR variables on overall liking pairwise and the probability weighted ranking was calculated according to the differences in wins and losses, denoted as pWinners [24].

2.3 Sum of ranking differences (SRD) method

SRD helps to determine the variables having significant impact on liking. In case of contradictory evaluation, the row-maximums are the natural choice of data fusion (defined as the combination of similarity rankings): The “best” values were selected as a reference, *i.e*. the golden standard unifies the best attributes of all methods used for ranking. The reference column can be considered as a hypothetical attribute composed from the best scoring of all evaluation methods. This means that the highest (row-)value of each attribute was inserted into the reference column. Comparison of the attributes was done using the sum of ranking differences (SRD) procedure [16]. The validation algorithms of the SRD method were published in 2011 [25]. For the SRD analysis, data is arranged in a matrix, where rows correspond to the JAR methods, while columns correspond to the JAR attributes. In our case higher values mean that the given method evaluates an attribute as more important. Hence, the row maximums have been chosen as the reference (benchmark) column of the SRD matrix (Max) in (Table 2).

Table 2
The input data matrix of the SRD after normalization (square root transformation of the original attributes). The reference columns contain the row maximum (Max) values.

	Size+	Size–	Color+	Color–	Flavor+	Flavor–	Salt+	Salt–	Thick/Thin+	Thick/Thin–	Stickiness+	Stickiness–	Max
OLS	3.2244	3.2300	3.3585	3.3516	3.4267	3.5246	3.2004	3.4741	3.3250	3.3372	3.5216	3.3081	3.5246
Penalty	3.1223	3.1936	3.2961	3.3440	3.3841	3.3029	3.3306	3.1409	3.2777	3.3267	3.3227	3.3732	3.3841
bPenalty	3.1221	3.1926	3.2962	3.3432	3.3846	3.3030	3.3319	3.1404	3.2775	3.3254	3.3219	3.3730	3.3846
GPCM	1.4399	3.1623	3.4641	3.8588	3.0077	3.8670	2.4498	3.0027	3.1560	2.8365	3.7433	1.3416	3.8670
PLS-dummy	3.1655	3.1731	3.1779	3.1772	3.1775	3.1879	3.1832	3.1885	3.1751	3.1763	3.1924	3.2441	3.2441
MLR	3.2430	3.1964	3.2690	3.3286	3.4647	3.5342	3.3175	3.3590	3.2241	3.1954	3.7087	4.3595	4.3595
wPAforJARMean	3.2946	3.2567	3.1767	3.1940	3.2319	3.1976	3.3185	3.2838	3.1753	3.2024	3.1957	3.1723	3.3185
wPAforGrandMean	3.1309	3.1901	3.1509	3.1479	3.1347	3.1211	3.2652	2.9887	3.1536	3.1476	3.1253	3.1563	3.2652

Variables having the highest impact on liking by each method are highlighted with grey. These values give the last, reference (benchmark) column

Normalization of the data is necessary due to the different scales of the investigated methods. We are looking for the attribute, which obtained the highest values from most of the JAR methods. After analyzing the JAR data using the above JAR data analysis methods, their results have been transformed using square root transformation. Several data preprocessing approaches were tried (*e.g.:* logarithmic transformation, standardization, *etc.*) and SRD analysis was run on every scaled data set. The consensus of the SRD runs showed that square root transformation is an acceptable choice, as no bias has been made by the transformation.

SRD is a novel, fast and entirely general method for the comparison of alternative solutions to the same problem – *e.g.* different measurement/calculation methods of the same property (in this case, JAR features). SRD is based on the comparison of the rankings produced by the different methods, *i.e.* the samples are ranked (in the order of magnitude) according to each method plus a reference method is also ranked. The differences between the rank numbers of each sample according to each method and the reference method are calculated, and these ranking differences are added up for each method. The reference method can be an exact "golden standard" or as in the present case the vector of row-maximums. Using the row-maximums as reference instead of the recommended experimentally determined sensory attributes is justified based on two main points: a) the maximum realizes the hypothetical "best" method extracting the largest average impact on overall consumer liking from each method; b) it is a well substantiated empirical finding that systematic errors of different laboratories (or methods) follow normal distribution. This is the base of proficiency testing (*e.g.* round-robin tests) [26]. Even if some biases remain, we are better off using row-maximums than any of the individual methods. The resulting values are called SRD values and the smaller they are, the closer the method is to the reference (in terms of ranking). These SRD values are usually normalized to enable the comparison of different SRD calculations:

$$SRDnorm = 100\mathrm{SRD}/\mathrm{SRDmax} \qquad (5)$$

where SRDmax is the maximum possible SRD value. The graphical representation helps to identify the significant attributes. This enables the user to evaluate the JAR attributes using the results of multiple methods. How the SRD values are calculated can be followed on an animation published as a supplement to our recent article on similarity metrics [27].

Table 3
Computation of the SRD values in the case of attribute Flavor–

	Max	rnk	Flavor–	rnk_6	Abs(diff_6)
OLS	3.52	6	3.5246	6	0
Penalty	3.38	4	3.3029	4	0
bPenalty	3.38	5	3.303	5	0
GPCM	3.87	7	3.867	8	1
PLS-dummy	3.24	1	3.1879	2	1
MLR	4.36	8	3.5342	7	1
wPAforJARMean	3.32	3	3.1976	3	0
wPAforGrandMean	3.27	2	3.1211	1	1
Sum					4

In Table 3, the Max values give the highest value for any attributes by the methods (the values of the variable having the highest impact on liking). After rank transformation, a variable (denoted by 'rnk') is created. The Flavor– variable contains the values given by each method for Flavor–. Signs at the end of the name of attributes indicate the too much (+) or the too low (**–)** region of the given attribute. The rnk$_6$ variable is the rank transformed Flavor– variable, while diff$_6$ is the rank difference of rnk and rnk$_6$. The sum of all elements in absolute diff$_6$ column gives an SRD$_6$ value, which in this case is equal to four. The same computation is done for all of the variables one-by-one.

Additionally, the SRD procedure not only calculates the sum of ranking differences but also contains two validation steps: 1) the randomization test gives features having a ranking different from random ranking and 2) we can assign uncertainty values to the SRD values with the help of leave-one-out (or eventually sevenfold) cross-validation.

Results of the JAR methods were calculated using R-project 3.1.0 [28].

Sum of ranking differences method was calculated with Microsoft Office Excel 2007 macro (available here):

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

3. Results and discussion

The input matrix (Section 3.1) and its transpose (Section 3.2) were subjected to SRD ranking to order and group the attributes and evaluation methods, respectively.

3.1. Ranking of the attributes

The maximum values have been inserted to the reference column of the SRD input matrix because the importance of a variable is determined based on its impact on liking. The theoretical SRD distribution was created and used as the number of rows were less than 14 (*n*=8).

The detailed results of the SRD computation are summarized in Table 4, and the graphical representation is shown in Fig. 1.

The SRD column of Table 4 contains the results of the sum of the rank differences for each attribute (SRD values). Column *p* % contains two probability values due to the discrete nature of the distribution and derived from the theoretical distribution of random ranking. The theoretically possible maximal SRD (MaxSRD) was computed for the given number of rows (8).

The last column of Table 4 contains the scaled SRD values between 0 and 100 (*SRDnorm*), which were calculated according to Eq. 5: MaxSRD=2k^2=32, because *n* is an even number (8) [25]. The additional rows of Table 4 mean the 5 % (XX1 – five percentiles), 25 % (Q1 – first quartile), 50 % (Med – median), 75 % (Q3 – last quartile) and 95 % (XX19). If the *SRDnorm* value of a given attribute is smaller than the average of the range for XX1 (5 % percentile), the JAR attribute is considered as significant at *p*=0.05 level. Based on these results, the following attributes were identified as significant: Flavor–, Stickiness+ and Color–.

Table 4
Ranking of JAR attributes and probability of random ranking

Ranking results		***p* %**		***MaxSRD*=32**
Name	**SRD**	**x < SRD**	**> =x**	***SRDnorm***
Flavor–	***4***	***0.02025***	***0.10007***	***12.5***
Stickiness+	***6***	***0.10329***	***0.38221***	***18.75***
Color–	***10***	***1.20908***	***3.07284***	***31.25***
XX1	12	3.12	6.83	
Color+	12	3.11759	6.83328	37.5
Flavor+	12	3.11759	6.83328	37.5
Stickiness–	14	6.91171	13.1449	43.75
Size–	18	22.448	34.3173	56.25
Salt–	18	22.448	34.3173	56.25
Thick/Thin+	18	22.448	34.3173	56.25
Q1	18	22.45	34.32	
Thick/Thin–	20	34.49	48.47	62.5
Size+	22	48.65	63.10	68.75
Med	22	48.65	63.10	
Salt+	24	63.25	76.64	75
Q3	24	63.25	76.64	
XX19	30	94.21	98.57	

Italic, boldface means that the attributes are significant at the *p*=0.05 level. The grey colored rows are the 5 % (XX1), 25 % (Q1), 50 % (Med), 75 % (Q3) and 95 % (XX19) percentiles.

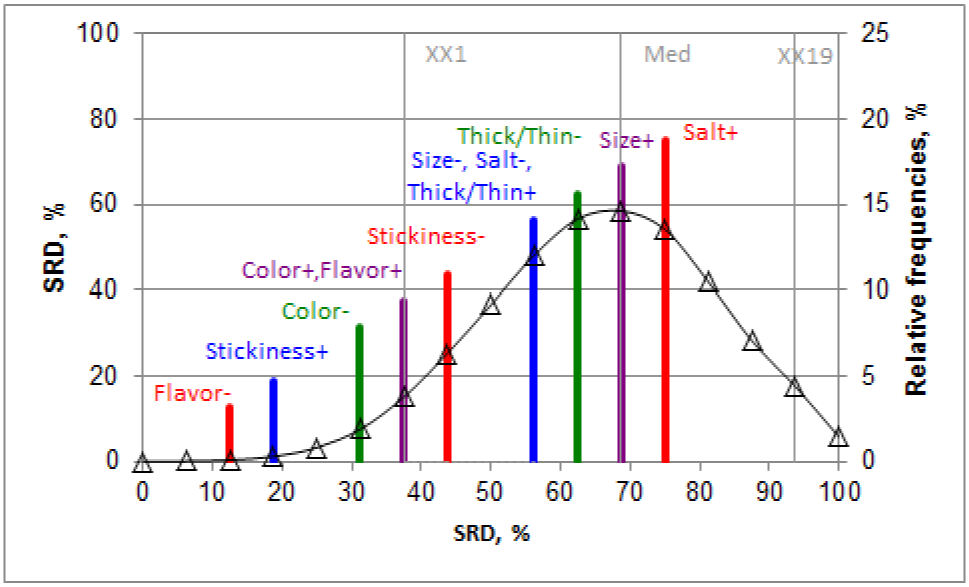

Figure 1 The scaled SRD values (between 0 and 100) of the attributes determined by sum of ranking differences. The row maximums were used as reference (benchmark) column. Scaled SRD values are plotted on *x* axis and left *y* axis, right *y* axis shows the relative frequencies where

triangles represent the exact counted values (black curve). The 5 % probability ranges (XX1), Median (Med), and 95 % (XX19) are also given.

Comparison is made based on the zero *SRDnorm* value, which means a hypothetical best method, *i.e.* there is no difference between the reference column (the maximal scoring) and the given attribute. The higher the value of *SRDnorm* is, the bigger the ranksum difference between the attribute and the reference column is. If the value of *SRDnorm* crosses the Gauss-curve say at *p*=0.10 then, the method ranks the variable as random with a 10 % chance (Figure 1).

Furthermore, the SRD method ranks the attributes based on their difference from the reference column (Max value). Hence, a ranking is made among the significant variables. The variable with the lowest *SRDnorm* value should have the largest effect on consumer liking and has to be changed first. Figure 1 gives the following order of attributes: Flavor–, Stickiness+ and Color–. The other attributes have higher *SRDnorm* values than the 5 % error limit (XX1). The evaluation of other attributes is indistinguishable from the evaluation of random numbers. It does not necessarily mean that their order carries no information. This simply means that they do not have significant impact on hedonic scores according to the best of the eight methods examined. High *SRDnorm* values can be reached if the attribute is ranked differently by the methods (consensus of the methods is low). The main advantage of this application of SRD method is that the results of multiple statistical methods give more reliable results and the main assumption of the SRD procedure corresponds to the maximum likelihood principle.

The SRD plot can be improved to make it more applicable in product development and JAR data analysis. On the mean drop plots of penalty analysis, the frequencies of consumers (*x*-axis) are plotted against the mean drops of the attributes of the products (*y*-axis). If high percentage of consumers rates an attribute as too much or not enough (consumer percentage), this attribute has significant impact on liking (high mean drop value). Then, the attribute will be located on the upper right quadrant. In Figure 2, percentages of consumers are plotted on the *y*-axis and the normed SRD % values are plotted on the *x*-axis. The main difference to the mean drop plot is that attributes having low SRD % values have significant impact on liking; hence, attributes located on the upper left corner are important for product development. Figure 2 helps to evaluate the consumers’ needs more precisely.

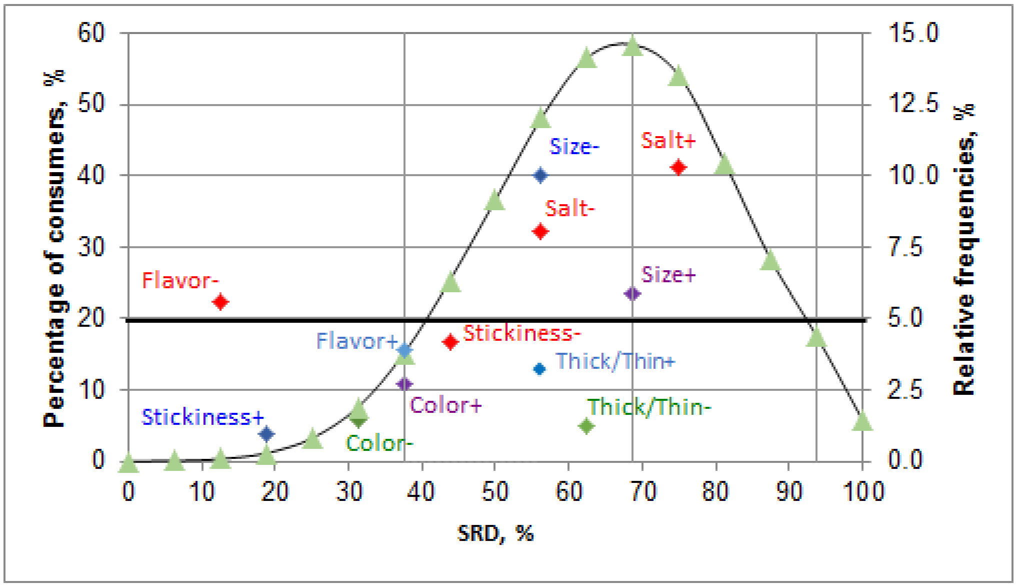

Figure 2 The combination of the scaled SRD values with the consumers' frequency values. The solid black line represents the 20 % threshold, which is generally applied in penalty analysis, as well.

Figure 2 is divided into two parts by the black line, which represents the 20 % threshold of the consumers. Those attributes, which were highlighted as true for the product by the consumers, are located above the line. Changing these attributes may result in great increase of liking for high number of consumers. The interpretation of the plot is similar to the SRD plot because the attributes, which have lower values than XX1 can be considered as significant at p = 0.05 level. Attributes located right from XX1 are non-significant according to the randomization test of SRD. The SRD plot gives not only a rank of importance but a comparison with random ranking. However, the new plot visualizes the attributes, which were highlighted as important by the consumers and the attributes, and those, which were significant by the SRD. The results show that Flavor– is the most important product attribute and the flavor should be strengthened to achieve better consumer acceptance. Stickiness+ and Color– were mentioned by only a lower percentage of consumers. These consumers disliked the product due to the too sticky and not enough intense color attributes. The consumers' evaluations were heterogeneous about size and saltiness; hence, the SRD analysis of the methods did not recognize them as significant. These attributes were important for the consumers but the attributes did not have great impact on their liking, which phenomenon is frequent in hedonic testing.

3.2 Ranking of the evaluation methods

SRD method can be used to assess not only the variables having the highest impact on consumer liking but the methods, enumerated in Table 1, as well. This way, several JAR data evaluating methods were compared and their consensus (or average) was used in SRD reference column. This approach is valuable in those cases when methods have no unambiguous results. It is supposed that all the applied methods evaluate the data with some error (bias + variance). In situations like this, the average result should be used because the random and systematic errors balance each other. In order to complete the comparison of the evaluation methods, all we need to do is to transpose the original input matrix (Table 2) and apply row-average instead of row-maximums. SRD gives the differences of the methods from their average results one-by-one. The zero point represents the average of all methods, the closest to zero is GPCM (Figure 3), while OLS has still significant results. The other methods (located over XX1) are overlapping with the distribution derived using solely random numbers for ranking.

Figure 3 shows characteristic groupings. MLR and PLS-dummy are located close to each other (having similar results), because both of them are multiple regression-based methods. Similar groupings are seen in the case of wPAfor-GrandMean and wPAfor-JARMean, while Penalty and bPenalty show no difference suggesting that bootstrapping do not improve penalty analysis.

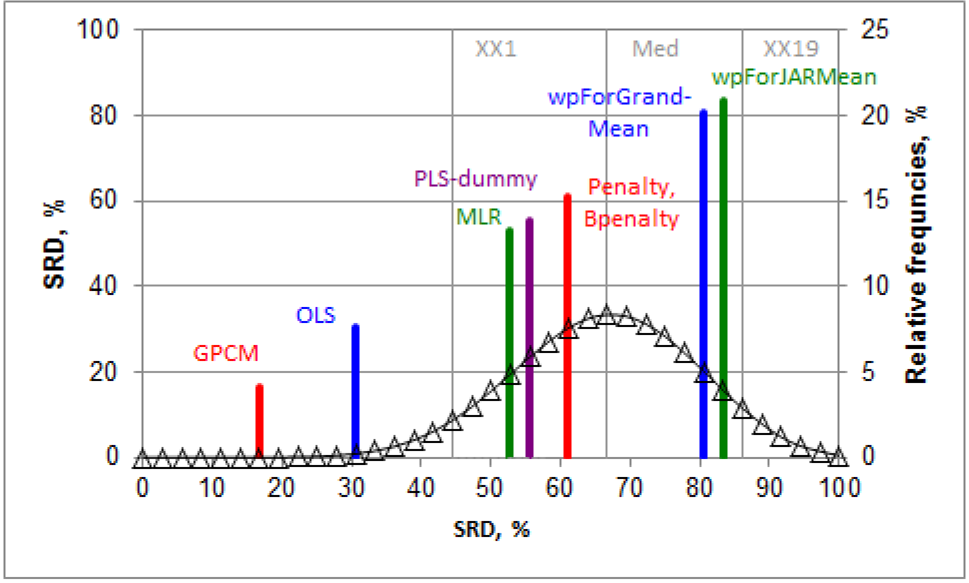

Figure 3 The scaled SRD values (between 0 and hundred) of the evaluation methods determined by sum of ranking differences. The row maximums were used as reference (benchmark) column. Scaled SRD values are plotted on *x* axis and left *y* axis, right *y* axis shows the relative frequencies where triangles represent the exact counted values (black curve). The 5 % probability ranges (XX1), Median (Med), and 95 % (XX19) are also given.

The SRD algorithm includes a leave-one-out cross-validation option, *i.e.* making the SRD procedure "*n*" times always on a smaller (*n*-1) data set and so rendering uncertainties to each feature SRD value. The created SRD values are plotted on a Box and Whiskers plot which gives the medians and quartiles. As shown by Fig. 4, the quartiles and min-max values of the clusters overlap which was confirmed by Sign test. This way, a similar but more sensitive result is given than in Figure 3. The Sign test gives the following five groups (starting from the lowest SRD): GPCM, OLS, MLR and PLS-dummy, Penalty and bPenalty and the final one is the group of wPAfor-GrandMean and wPAfor-JARMean.

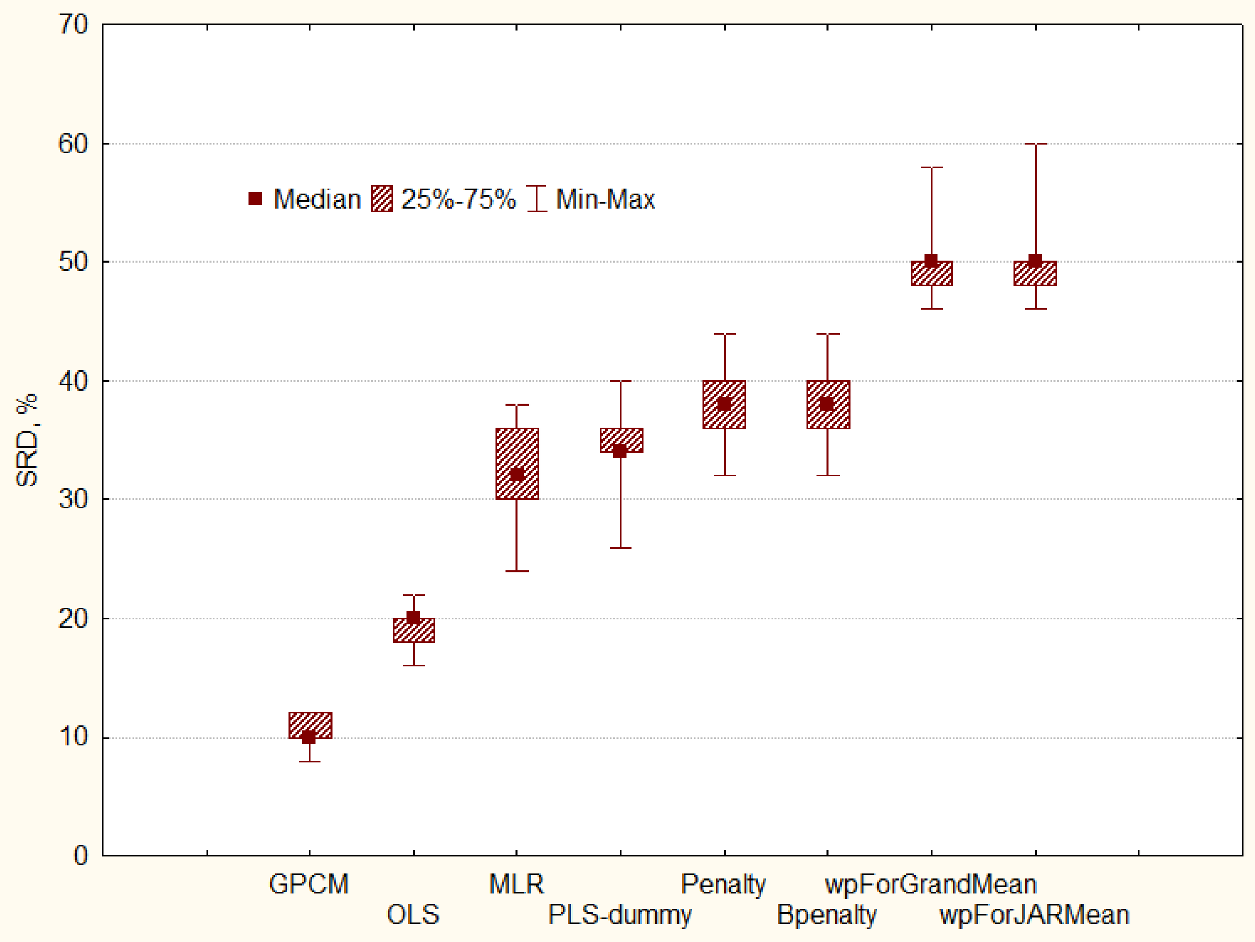

Figure 4 Box and whiskers plot of the SRD values after leave-one-out cross-validation

4. Conclusions

There are several different methods published in the literature for just-about-right (JAR) data analysis; we looked for the optimal combination of JAR methods, which split the JAR scale and takes into account the hedonic ratings as well. However, these methods may give different, sometimes contradictory results. We created a solution, which gives a more reliable variable selection for JAR attributes. The reliability is supported with exact statistical tests and derived theoretical probabilities. The proposed method is not a new JAR data evaluating method but an approach, which helps to solve problems, when the results of general JAR data evaluation methods deviate (as usually happens). Sum of ranking differences (SRD) method has successfully and frequently been applied in chemometrics earlier. Due to the different data sets or variables measured on different scales, scaling of data is necessary. The SRD method proved to be a successful tool to rank the JAR variables based on their differences from the reference column (maximum values of the methods) and to prove their significance. This enables to create ranking between the product attributes. The SRD method gives a recommendation for how to optimize the products based on the results of several JAR methods. It can easily be determined, which attribute and which direction of the attribute should be changed to improve the products and reach higher consumer liking scores. Further advantages of the method are that the set of the included methods can be easily changed. The results can be interpreted easily, there is a freely accessible macro to run SRD and it is user friendly.

If the frequency data of the consumers is integrated into the original SRD plot the significant and important (for the consumers) product attributes can be identified. Generally, the practical importance of the provided methodology is that the results of multiple methods give a more reliable outline about the products. Multicriteria optimizations apply weights, but no such subjective factors (weights) should be introduced in the SRD methodology. Furthermore, it enables to focus on the most important key attributes during product development.

Due to the characteristics of SRD, analysis of the transposed matrix gives also valuable information. In our case, the rank of the applied methods gave that GPCM gives the most “average” results. Hence, GPCM is suggested instead of the other analyzed methods. This feature of SRD is useful in several fields *e.g.* the performance of model validation parameters when evaluating QSAR and binary QSAR models [29].

Acknowledgement

The authors greatly thank to Lori Rothman for providing the ASTM data set. KH thanks the support of OTKA, contract No K119269. Supported By The Únkp-16-4 New National Excellence Program Of The Ministry Of Human Capacities.

References

[1] ASTM E253-15, ASTM E253. Terminology relating to sensory evaluation of materials and products, (2015). doi:10.1520/E0253-15

[2] ASTM E456, ASTM E456, Standard Terminology Relating to Quality and Statistics, (2013). doi:10.1520/E0456

[3] L. Rothman, M.J. Parker, ASTM MNL63; Just about Right (JAR) Scales: Design, Usage, Benefits, and Risks, West Conshohocken, 2009. doi:10.1520/MNL63-EB

[4] D.J. Best, J.C.W. Rayner, D. Allingham, A note on formulae for CMH statistics for JAR data, Food Qual. Prefer. 31 (2014) 19–21. doi: 10.1016/j.foodqual.2013.07.006

[5] J.F. Meullenet, R. Xiong, C. Findlay, Multivariate and Probabilistic Analyses of Sensory Science Problems, 1st ed., Wiley-Blackwell, Ames, Iowa, 2007.

[6] C.D. Goerlitz, J.F. Delwiche, Impact of label information on consumer assessment of soy-enhanced tomato juice, J. Food Sci. 69 (2004) S376–S379. doi:10.1111/j.1365-2621.2004.tb09952.x

[7] M. Gacula, S. Rutenbeck, L. Pollack, A.V.A. Resurreccion, H.R. Moskowitz, The Just-About-Right intensity scale: Functional analyses and relation to hedonics, J. Sens. Stud. 22 (2007) 194–211. doi:10.1111/j.1745-459X.2007.00102.x

[8] D. Plaehn, J. Horne, A regression-based approach for testing significance of “just-about-right” variable penalties, Food Qual. Prefer. 19 (2008) 21–32. doi:10.1016/j.foodqual.2007.06.003

[9] R. Xiong, J.F. Meullenet, A PLS dummy variable approach to assess the impact of jar attributes on liking, Food Qual. Prefer. 17 (2006) 188–198.

doi:10.1016/j.foodqual.2005.03.006

[10] K. Heberger, R. Rajko, Variable selection using pair-correlation method. Environmental applications, Sar Qsar Environ. Res. 13 (2002) 541–554.
doi:10.1080/10629360290023368

[11] A. Gere, S. Kovacs, K. Pasztor-Huszar, Z. Kokai, L. Sipos, Comparison of preference mapping methods: a case study on flavored kefirs, J. Chemom. 28 (2014) 293–300. doi:10.1002/cem.2594

[12] U. Gromping, Variable importance assessment in regression: Linear regression versus random forest. Am. Stat. 63 (2009) 308–319.
doi: 10.1198/tast.2009.08199

[13] L. Breiman, Random Forests, Mach. Learn. 45 (2001) 5–32.
doi: 10.1023/A:1010933404324

[14] J. W. Johnson, A heuristic method for estimating the relative weight of predictor variables in multiple regression. Multivariate Behav. Res. 35 (2000) 1–19.
doi: 10.1207/S15327906MBR3501_1

[15] J. Bi, J. Chung, Identification of drivers of overall liking – determination of relative importances of regressor variables, J. Sens. Stud. 26 (2011) 245–254.
doi:10.1111/j.1745-459X.2011.00340.x

[16] K. Heberger, Sum of ranking differences compares methods or models fairly, TrAC Trends Anal. Chem. 29 (2010) 101–109.
doi:10.1016/j.trac.2009.09.009

[17] L. Sipos, Z. Kovacs, D. Szollosi, Z. Kokai, I. Dalmadi, A. Fekete, Comparison of novel sensory panel performance evaluation techniques with e-nose analysis integration, J. Chemom. 25 (2011) 275–286.
doi:10.1002/cem.1391

[18] K. Kollar-Hunek, K. Heberger, Method and model comparison by sum of ranking differences in cases of repeated observations (ties), Chemom. Intell. Lab. Syst. 127 (2013) 139–146.
doi:10.1016/j.chemolab.2013.06.007

[19] A. A. Gowen, G. Downey, C. Esquerre and C. P. O'Donnell, Preventing over-fitting in PLS calibration models of near-infrared (NIR) spectroscopy data using regression coefficients, J. Chem. 25 (2011) 375–381.
doi: 10.1002/cem.1349

[20] B. Skrbic, K. Heberger, N. Durisic-Mladenovic, Comparison of multianalyte proficiency test results by sum of ranking differences, principal component analysis, and hierarchical cluster analysis., Anal. Bioanal. Chem. 405 (2013) 8363–75.
doi:10.1007/s00216-013-7206-5

[21] A. Racz, D. Bajusz, K. Heberger, Consistency of QSAR models: Correct split of training and test sets, ranking of models and performance parameters., SAR QSAR Environ. Res. 26 (2015) 683–700.
doi:10.1080/1062936X.2015.1084647

[22] J.A. Koziol, Sums of ranking differences and inversion numbers for method discrimination, J. Chemom. 27 (2013) 165–169.
doi:10.1002/cem.2504

[23] K. Heberger, R. Rajko, Generalization of pair correlation method (PCM) for nonparametric variable selection, J. Chemom. 16 (2002) 436–443.

doi:10.1002/cem.748

[24] A. Gere, L. Sipos, K. Heberger, Generalized Pairwise Correlation and method comparison: Impact assessment for JAR attributes on overall liking. Food Qual. Prefer. 43 (2015) 88–96. doi:10.1016/j.foodqual.2015.02.017

[25] K. Heberger, K. Kollar-Hunek, Sum of ranking differences for method discrimination and its validation: comparison of ranks with random numbers, J. Chemom. 25 (2011) 151–158. doi:10.1002/cem.1320

[26] W.J. Youden, Statistical Manual of the Association of Official Analytical Chemists: Statistical Techniques for Collaborative Test. AOAC International, Gaithersburg (1975)

[27] D. Bajusz, A. Racz, K. Heberger, Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J. Cheminform. 7 (2015), p. 20. doi: 10.1186/s13321-015-0069-3

[28] R Development Core Team, R: A language and environment for statistical computing, (2016). https://www.r-project.org/.

[29] N.S. H. N. Moorthy, N.M.F.S.A. Cerqueira, M. J. Ramos, P. A. Fernandes, Ligand based analysis on HMG-CoA reductase inhibitors. Chemometr. Intell. Lab. 140 102-116. doi:10.1016/j.chemolab.2014.11.009