Biological Agriculture and Horticulture, 2010, Vol. 27, pp. 25–40

Standardization and Validation of the Visual Evaluation of Biocrystallizations

Machteld Huber[1,]*, Jens-Otto Andersen[2], Johannes Kahl[3], Nicolaas Busscher[3], Paul Doesburg[1], Gaby Mergardt[3], Sebastian Kretschmer[4], Aneta Zalecka[4], Aumaporn Meelursarn[4], Angelika Ploeger[3], Dré Nierop[5], Lucy van de Vijver[1] and Erik Baars[1] (2010)

[1] Department of Healthcare and Nutrition, Louis Bolk Institute,Hoofdstraat 24, NL-3972 LA Driebergen, The Netherlands. [2] Biodynamic Research Association, Denmark. [3] Department of Organic Food Quality and Food Culture, University of Kassel, Germany. [4] Department of Organic Food Quality and Food Culture, University of Kassel, Germany. [5] Muvara Statistics, The Netherlands

ABSTRACT

The biocrystallization method has been used as a holistic method for evaluating food samples since the 1930s and has, as recent reviews have concluded, been successful in discriminating farming systems in several studies. However, it is noted that a standardization and validation for more objective means of evaluation of biocrystallization pictures and a clearly defined and communicable language, as well as possibilities for verification, are still lacking. During the years 2002–2006 a project for standardizing and validating the visual evaluation of these biocrystallizations took place, performed by a group consisting of three European research institutes. As no pre-set methodology nor norms for the visual evaluation existed, the choice was made to work according to existing norms on sensory analysis and to adapt these to the visual evaluation. The main norm, ISO-Norm 11035, for establishing sensory profiles was adapted towards these pictures. A panel was formed, mainly morphological descriptors were selected and defined, a scale with references was established, the panel trained and tested for pictures, produced on the basis of carrots of different qualities. Reliability and validity aspects of the panel, as well as the structure of the set of criteria were evaluated. According to standard rates for diverse statistical tests the panel has been validated for the evaluation of carrots with a defined set of descriptors. The set of descriptors contains sub-domains which can be used for further development of the evaluation method, towards a more integrated, holistic ‘Gestalt’-evaluation, and towards finally interpretation of these pictures. The followed process and results are unique in connection to this method.

*Corresponding author – m.huber@louisbolk.nl

INTRODUCTION

The biocrystallization method, or copper chloride ($CuCl_2$) crystallization method, was developed by Pfeiffer (1930) as a holistic method. Since then, it has been used for evaluating food samples, mainly in organic agricultural studies (Engquist, 1975) and as a holistic medical diagnostic instrument, mainly by anthroposophic doctors (Bessenich, 1951; Selawry & Selawry, 1957).

The evaluation of biocrystallization pictures is commonly performed by means of visual evaluation. In the last decade computerized image analysis has been added (Andersen *et al*., 1999), which is still in further development. Until now visual evaluation is superior in discriminating differences between pictures and in identifying pictures as originating from different cultivation systems (Kahl, 2007). In fact the visual evaluation directs the development of the computerized evaluation. In the history of the visual evaluation of biocrystallization pictures, several well trained specialists have been shown to be capable of evaluating crystallization pictures in such a way that they could identify them correctly as coming from different farming systems or different processing techniques (Balzer-Graf, 1991, 1994, 1996, 2001; Mäder *et al*., 1993; Alföldi *et al.*, 1996; Granstedt & Kjellenberg, 1997; Weibel *et al.*, 2001; Andersen, 2001). Although several authors have described their methods of evaluation, among whom Schudel *et al.* (1980a, 1980b) and Amons *et al*. (1990) were the first to work with a panel, until now no consensus between the different experts has been reached (Selawry & Selawry, 1957; Von Hahn, 1962; Petterson, 1967; Engquist, 1970; Schudel, 1980a, 1980b; Balzer-Graf, 1994, 1996; Amons, 1999). In the evaluations of specialists in visual evaluation a hierarchy of five levels of judgements can be discriminated. These levels are

(1) Quantifiable judgements about single morphological and local features (e.g. length of sideneedles), (2) Qualitative descriptive judgements, connected to single morphological features (e.g. regularity of ramifications), (3) Qualitative descriptive judgements of a higher order, that describe gestures in the whole picture (e.g. coordination, integration of the features), (4) Qualitative interpretative judgements about plant-physiological aspects (e.g. ripeness, root-typical), based on references from defined samples), 'Gestalt-judgements', (5) Qualitative interpretative judgements of the highest order, namely on the food quality of a product, relating to clearly defined concepts of quality.In the available literature it is not always possible to retrace the features in the pictures on which certain quality judgements were based.

Recent reviews on comparative studies of the effects of different farming systems on crop and product quality have concluded that the biocrystallization method has in several studies successfully discriminated the farming systems, but it is noted that various methodological difficulties and insufficiencies exist, including a lack of a validated methodology and the need for more objective means of evaluation of the pictures. A clearly defined and communicable language, as well as possibilities for verification, is still lacking (Woese *et al.*, 1995; Soil Association, 2001). To overcome these problems, three European institutes, Louis Bolk Institute (LBI) in The Netherlands, University of Kassel, Department of Organic Food Quality and Food Culture (Uni-Kassel) in Germany, and Biodynamic Research Association Denmark (BRAD), set up the goal to develop a standardized laboratory procedure, including a communicable method of visual evaluation of crystallization pictures, that would meet scientifically accredited principles and could reach standardization and a first step on the route towards a validation of a judgement panel. A documented and characterized methodology including crystallization chamber, laboratory procedures and image analysis procedures for quantification of the crystallization pictures has been developed or is under development (Kahl, 2007).

The present study was performed to reach a standardized, validated, scientifically acknowledged and communicable method of visual evaluation of crystallization pictures during the years 2002–2006. In the study, the questions that were evaluated are categorized in three groups: (A) The reliability of the panel and its development (standardization), (B) The validity of the panel in discriminating products significantly (validation) and (C) Evaluation of the structure of the set of criteria used in assessments.

MATERIALS AND METHODS

Sensory norm adapted towards visual evaluation

Currently no pre-set methodology nor norms for the standardization and validation of visual evaluation exist. The method used in this study was based on previous experience developed for panel-judgement of blood crystallization pictures of the Louis Bolk Insitute (Amons *et al*., 1999), and now connected to the scientific methodology of sensory analysis. Because of the parallelism between evaluation of pictures and the evaluation of taste, both based on the use of human sense experiences, the choice was made to work according to existing norms on sensory analysis and adapt these to the visual evaluation.

The main norm on sensory analysis is ISO-Norm 11035, 1994 (E) – Sensory Analysis – identification and selection of descriptors for establishing a sensory profile by a multidimensional approach (ISO, 1994). This norm clearly describes the steps to be taken towards a standardized panel for sensory profiles. This norm was adapted for use in the standardization of visual evaluation, by modifications from food samples to crystallization pictures (Kretschmer, 2003). Adaptations were made with respect to specific food-connected terms, e.g. ‘product’ was adapted to ‘picture’, etc. Next to the ISO-Norm 11035, other connected ISO-Norms were used for specific items (5492-1992; 6658

1985; 6564-1985; for details see Huber *et al.*, 2007). In addition, tests were performed to study the validity of the test in discriminating between different types of food samples.

According to ISO-Norm 11035 the following steps for developing the evaluation instrument were taken: formation of the panel, identification of the largest possible number of descriptive terms, preliminary sorting of descriptors, reduction of descriptors, finally achieving a list of maximum 15 descriptors, choice of references, training of the panel to use this reduced list and performing panel tests. Two adaptations were made towards the norm: the descriptors chosen were not always mono-dimensional, which would be criteria on level 1 above. Using just criteria of level 1 was considered too inefficient, as only a very minor part of the visual information of the pictures is described by this. The choice was made for quantitative and qualitative descriptive morphological criteria, as well as qualitative descriptive criteria of a higher order (levels 1, 2 and 3). Secondly, principal component analyses for further grouping and reducing of criteria was performed afterwards and not during the process. This was because of the aim to train a panel with a refined discriminative skill, as a basis for future more complex judgements.

Developing the evaluation instrument

The norm states a minimum of six persons in a panel. A panel was formed from members of the project group, both experienced and inexperienced at the start. During the project altogether nine persons participated. A core group of six persons, including the panel leader, participated in all panel meetings and tests during the project.

The standardization steps were worked out with pictures produced on the basis of carrots. The crystallization pictures originated from archive material of carrot studies performed on samples from different connected studies by the University of Kassel (Kahl, 2007) and a microwave heating study of BRAD (Andersen *et al.*, 2005). These series contained mainly five to six replicate pictures per sample preparation, and repeated sample preparations. As the panel would be using photographs for the evaluation, crystallization pictures were photographed. For this an Olympus E300 digital camera, fitted with a 14–45 mm Zuiko Digital lens was used. Camera software version 1.0, with record mode HQ (jpeg file format 1/8 compression), aperture priority F10.0, centre weighed average metering, sharpness +1; saturation +1; contrast +2, white balance 4500K, ISO 200, exposure compensation +0.3EV and focusing mode single AF. Illumination of the crystallizations was by means of dark field illumination. For this a wooden lightbox (dimensions 390 × 390 × 160 mm l × w × h) was created. Two 8 Watt/33 cool white TL-tubes were positioned 45 mm from the sides of the lightbox (contact points at 55 mm height). Two 390

× 60 mm (l × h) pieces of multiplex (18 mm thick) were positioned 90 mm from the two sides of the lightbox (155 mm apart from each other); providing the dark field illumination. All the wood was painted matt black to increase the illumination effect. The lid consised of a 4 mm thick glass plate covered with matt black foil, with a centrally positioned 100 mm diameter opening in the foil. Crystallization plates were placed over this spacing, and covered with a 90 mm diameter black mask. From a series of pictures, those were removed which had technical failures, e.g. were multi-centred, and/or had an extreme evaporation time, as these differ markedly morphologically from the pictures around the mean crystallization time. And if a series had a strong unstable presentation of features among the replicates, which sometimes occurred, as variation is not yet completely ruled out in the method (Kahl, 2007), it was removed too. The remaining pictures were coded per series by an external person. Then, blindly, per series, three numbers of pictures per sample were randomly selected to be evaluated. The pictures were multiplied for judgement and were only decoded after testing.

Visual evaluation of crystallization pictures is generally performed by assessing first the impression of the typical character or ‘Gestalt’ unity, which the whole picture brings forward, followed by discriminating the different underlying morphological features. The panel group studied at the start a wide variety of pictures and identified the largest possible number of descriptive terms. The descriptors were reduced in several discussion rounds to a list of 14 descriptors or criteria. The norm states a maximum of 15 descriptors. The meaning of the criteria was defined and agreed upon by consensus.

An ordinal scale of intensity, ranging from 1 to 9, was chosen to rank the criteria, so also the more qualitative and complex criteria were quantified. For all criteria, reference pictures were chosen which connected to the scale intensities 1-4-7-9. During the training period the definitions were in a process of continuous refinement. A process of training started to constitute evaluation profiles, using the scale. The group met 20 times in the period 2002-2006 in the sensory laboratory at University of Kassel (designed according to ISO 8595, 1988 – about requirements for sensory cabins), for training and performing tests. During these rounds, panel members were situated in individual sensory booths, while the panel leader presented the pictures one after another, with a clean scoring form. Evaluations lasted a maximum of 15 min per picture. Following this, training took place via email and telephone conferences. All tests were preceded by new training rounds. Towards the end of the project, the criteria and definitions were fixed in the state they were to be used in the final testing. The training process, according to the norms, is summarized in Figure 1.

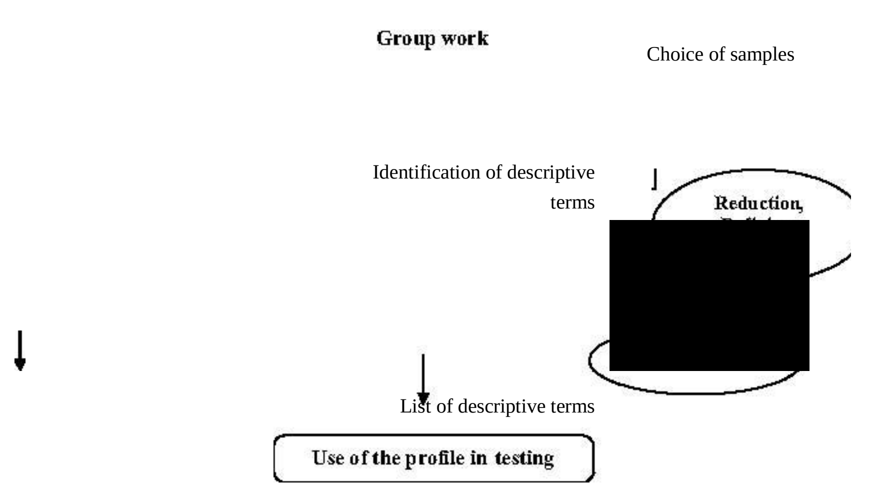

FIGURE 1. Summarizing the process of the norms.

The tests performed

Carrot pictures were visually evaluated in six panel tests, performed on different types (or qualities) of carrots, with varying questions (Table 1). A test with pictures of carrots, treated by microwave heating, was included as to include extreme scores of the used criteria and to experiment with extra criteria, deriving from these special pictures. Additional training of the panel was needed for this evaluation. In four of these series (7100, 4100, 6500,

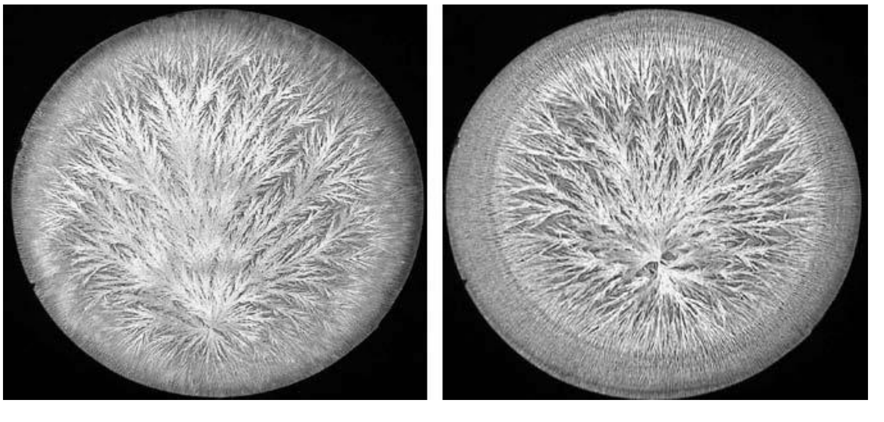

FIGURE 2. Crystallization pictures of a fresh (left) and a freeze dried (right) carrot

TABLE 1 The tests performed from 2003–2005 and number of panel members per test.

Series	Samples Compared	Source of pictures	Date of testing	Panel N=
7100	Fresh and freeze-dried carrots	Kassel experimental pictures	October 2003	7
4100	Fresh and freeze-dried carrots	Same Kassel experimental pictures as 7100 series	November 2004	9
Microwave	Fresh and microwave heated (MH) carrots	BRAD experiment	February 2005	8
4500	Carrots, (a) different varieties and (b) fertilizer levels 2004	Kassel – Fleck, BÖL	April 2005	7
6500	Carrots, (a) different varieties and (b) fertilizer levels 2004	Kassel – Fleck, BÖL Partly the same as 4500 series	June 2005	8
1100	Carrots, (a) different varieties and (b) production systems (organic and conventional)	Kassel – Market samples BÖL 2004	November 2005	8

1100), pictures were evaluated three times, over three days. In series 4500 and the microwave heating (MH) series no triplicate pictures were used, so no intra-observer variability could be assessed. A set of pictures was evaluated on one day and differently coded presented on the next days, arranged in a random order. Examples of samples of fresh and freeze-dried carrots are shown in Figure 2 and of fresh and microwave heated carrots, in an adapted mixing ratio of $CuCl_2$ and substance, in Figure 3.

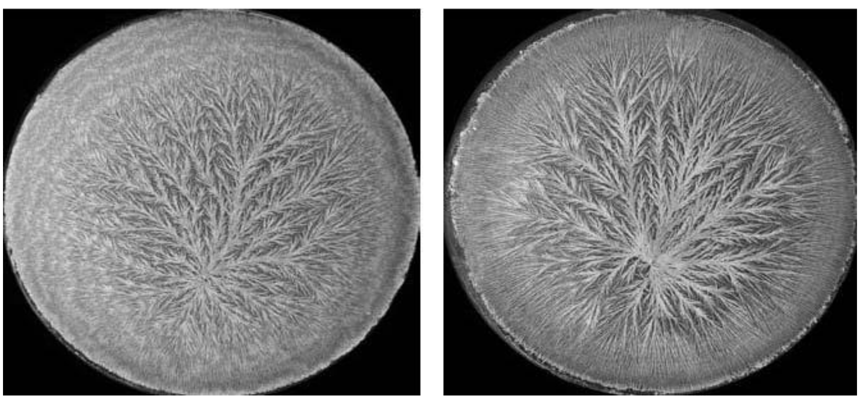

FIGURE 3. Crystallization pictures of two carrot varieties: Rodelika (left) and Rothild (right).

Statistics

SPSS statistical software (version 13.0) and MATLAB software (version 7.1.0) were used to perform statistical analyses. The reliability of the panel and its development was evaluated by the intra-observer and inter-observer variability of the panel and its improvement. The intra-observer variability of the individual evaluators was assessed by determining the standard deviation within repeated measures through calculating the deviations from the mean over three repeated measurements for each picture. An SD of zero would be ideal. The lower the standard deviation, the better the repeatability is considered. Additionally the one way single measure Intraclass Correlation Coefficient or Reliability Test (ICC, Shrout & Fleiss, 1979) was computed as a relative measure for the consistency of the repeated measures per person. Ideally the ICC has a value one.

In the perspective of the aim of standardization of the panel a standard rate for reliability is of importance. In the literature, different values for a minimum acceptable level of reliability can be found, either reliability being acceptable if larger than 0.5 (Streiner & Norman, 1995), or moderate reliability with measures between 0.40 and 0.59, substantial reliability if between 0.60 and 0.79, and outstanding reliability if 0.80 and higher (Landis & Koch, 1977a, 1977b).

The inter-observer variability was assessed by calculating for all evaluators the mean difference from the median and the standard error from the median, for the total set of pictures. The standard error was calculated as the root mean squared error of the differences from the median. A rate for standard error from the median is not given. It is of importance to observe whether or not there has been improvement.

The validity of the panel in discriminating products significantly, was evaluated by using independent t-tests (two-tailed) and discriminant analysis. A p-value < 0.05 is considered statistically significant. In the MH-test a paired t-test was applied. To test the discriminating skill of the panel, discriminant analysis was applied, resulting in cross-validated (CV) correct rates of classification. This CV correct rate of classification is robust and not sensitive to overfitting. Concerning a standard rate for discriminatory capacity: As a rule of thumb a correct rate above 70% can be considered a significant difference between two groups, while 50% is the lowest value indicating a random assignment to two groups (Bijlsma *et al*., 2006).

Evaluation of the structure of the set of criteria used in assessments was performed by principal component analysis to investigate if criteria occurred clustered and to identify sub-domains of criteria. Next to this Cronbach's α was calculated (except of the MH-test) to evaluate the internal consistency of the set of criteria. Cronbach's alpha: $x > 0.8$ = good; $0.6 < x > 0.8$ = moderate; $x < 0.6$ = bad (Heus *et al.*, 2002).

RESULTS

As a result of the described group process, 14 descriptive criteria, in a mixture of English and German language (Table 2), were defined and agreed upon in consensus. These criteria were connected to an ordinal scale of 1 to 9. Criteria that represent a hypothesized ‘better quality’ characteristic, score a higher number if more present, whereas criteria that represent a supposedly ‘weak quality’, score higher when absent. The same descriptors can be applied to pictures from several products. The questions that were evaluated with these criteria and scale were (A) The reliability of the panel and its development, (B) the validity of the panel in discriminating products significantly and (C) the evaluation of the structure of the set of criteria used in assessments.

TABLE 2

Overview of the 14 descriptive criteria as used in panel judgements,
connected to a scale of 1 to 9.

1 Integration 2 Centre
co-ordination
3 Durchstrahlung 4
Beweglichkeit 5 Regulation of
ramifications 6 Clear stems
formed 7 Fullness with
sideneedles
8 Thinning out (considered undesirable, so higher if absent) 9
Length of sideneedles 10 Substance spirals (undesirable, higher
if absent)
11 Dense radial formations and/or Flechtwerke (undesirable, higher if absent)
12 Lemniscate forms (undesirable, higher if absent) 13
Curly needles (undesirable, higher if absent)
14 Quernadeln (undesirable, higher if absent)

The reliability of the panel and its development

Intra-observer variability

The inter-observer variability of this panel for the repeated measures was low and improved slightly during the progress of the project. The mean SD for the panel was 0.47 (range 0.37–0.64) for series 7100 and 0.45 (range 0.35–0.64) for series 4100, 0.41 (range 0.37–0.49) for series 6500, 0.41 (range 0.30–0.53) for series 1100. So over the successive series, the mean SD’s were low and decreased during time from 0.47 → 0.45 → 0.41 → 0.41, thus showing an improvement. The mean ICC for the series 7100, 4100, 6500 and 1100 were respectively 0.68 (range 0.58–0.79); 0.67 (range 0.38–0.8); 0.68 (range 0.51–0.84) and 0.58 (range 0.39–0.82). These are all higher than 0.50 and can be classified as mostly substantial (> 0.60) or as moderate reliable (0.58).

Inter-observer variability

The development of the mean outcome of the difference from the median was -0.11 → –0.16 → –0.13 (MH-test) → –0.13 → –0.05 → –0.04, which is an improvement. Concerning the standard error from the median (root mean squared error), the standard error outcomes chronologically were 0.91 → 0.89 → 1.41 (MH-test) → 0.81 → 0.75 → 0.77, which overall is an improvement.

The validity of the panel in discriminating products significantly

Concerning the results of the independent t-tests, of all the series, out of eight sum scores four were significantly discriminating (= 50%). If the separate criteria are taken into account, in the eight series with 14 criteria, giving 112 possibilities to discriminate, 81 (72%) criteria were significantly discriminating.
The MH-test showed that the panel is able, with a minimum of extra training, to adopt some new criteria and to evaluate new types of pictures. Example of the results are presented graphically in Figures 4 and 5.
In discriminant analysis the results of the CV correct rates were 90% → 94% → 59% → 85% (MH-test, question 1) → 73% (MH-test, question 2) → 98% → 71% → 76% → 86%, which are all, except one, stating a significant

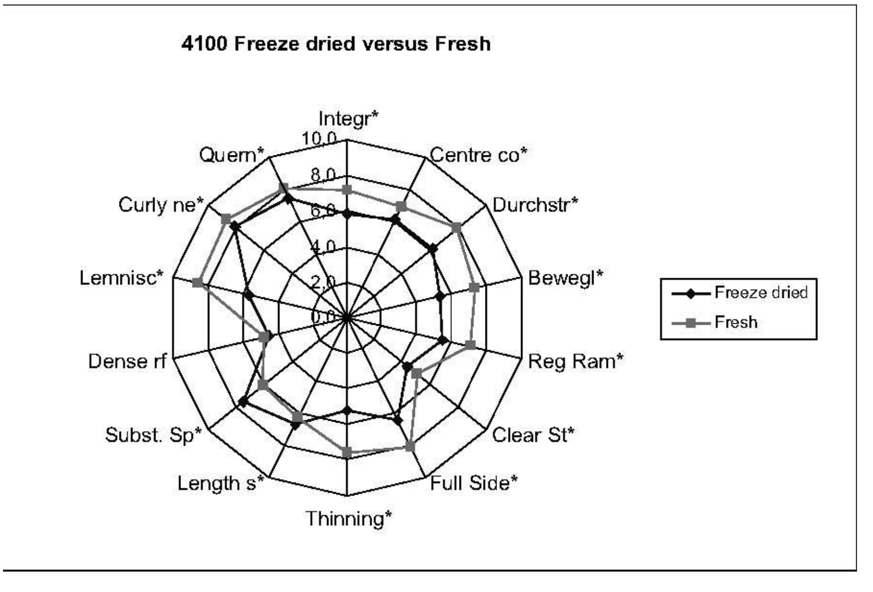

FIGURE 4. Graphical presentation of the mean panel scores, with criteria discriminating significantly (*) between fresh and freeze dried carrots, in test 4100.

difference, as being above 70%. The discriminant analysis also investigated the question if the panel results would improve, if, per series, one evaluator would be left out. The conclusion is that this hardly influenced the results.

Evaluation of the structure of the set of criteria used in assessments

From the PCA four sub-domains could be distinguished, each containing one or more criteria:
(1) ‘Integration’, ‘Centre coordination’, ‘Clear stems formed’ and (the absence of) ‘Dense radial formations’, (2) ‘Durchstrahlung’, (3) ‘Length of sideneedles’ and (the absence of) ‘Substance spirals’, and (4) ‘Beweglichkeit’, ‘Regularity of ramifications’, ‘Fullness with sideneedles’, ‘Thinning out’, ‘Lemniscate forms’, ‘Curly needles’ and ‘Quernadeln’. When analysing the discriminating power of these sub-domains, even better results are obtained.
Cronbach’s α was calculated, for the whole set of criteria, as well as for the sub-domains. The internal consistency of the criteria was bad (below 0.6) in three out of four series, if the whole set was calculated. When calculated for sub-domains 1, 3 and 4, then 10 out of 12 times the set of criteria shows a moderate to good correlation (above 0.6 or above 0.8). When Cronbach’s α for sub-domains 1, 3 and 4 was calculated for all series as a whole, all sub-domain sets of criteria showed an almost good correlation (almost higher than 0.8).

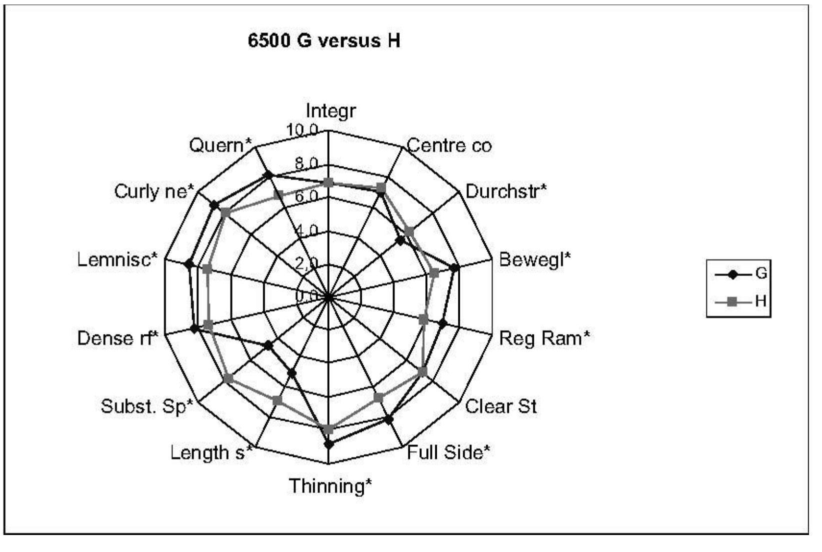

FIGURE 5. Graphical presentation of the mean panel scores, with criteria discriminating significantly (*) between two varieties of carrots, in test 6500. G = Rodelika, H = Rothild.

DISCUSSION AND CONCLUSION

The described process was undertaken with the aim of standardizing and validating a panel for visual evaluation of biocrystallization pictures and with this meet the criticism that, however successful individual experts have been in evaluations in organic research, the method lacked until now a standardized methodology and objective means of evaluation. With the described process, a tool has been developed for standardized visual evaluation of this laboratory method. Criteria have been developed on carrots, but can be used for different products, according to the experience of the authors. These criteria have been defined and references are available in connection to a scale. All this is performed by consensus, according to an adapted ISO-Norm for sensory evaluation.

A panel was trained and panel tests for establishing evaluation profiles of crystallization pictures were successfully performed. With this standardized visual evaluation method an essential step has been taken in the development of the biocrystallization method as a tool for quality research in organic and conventional farming.

For this standardization process, the sensory ISO-Norm 11035 was used. This norm appeared very well applicable for this other area of use of the human senses. A difference with usual sensory analysis on products is that, with each new batch of products, in sensory analysis the scale is defined anew, based on the range in the presented products. In the biocrystallization evaluation this was considered to be problematic. Standardization was at the moment the first aim and the achievement is that a list of clearly defined criteria and references is available.

Two concessions have been made to the Norm 11035. One was not to use just one-dimensional descriptors, as only very little of information in the pictures is covered by those. It was argued that it would be too time-consuming, first to perform a standardization on one-dimensional descriptors and then repeat the process with more complex ones. The challenge was to make more complex descriptors quantifiable, which was performed successfully, according to the results of this study. A second decision was not to perform a PCA on the descriptors in an early stage, to reduce the number of these, but to do this later, once a reliable panel was formed. One argument for this was to train a panel with a refined discriminative skill, as a basis for future more complex judgements, which might be lost when only a few descriptors would be left over from an early stage. The second argument was that the panel was changing in composition during the process and an unstable panel might not lead to a reliable information.

A weakness in the process is that from repeated series of one sample, a series was removed if it had a strong unstable presentation of features among the replicates, which sometimes occured. This is a problem of variation within

the method that is still not solved. It was however considered not to be a problem to be solved by the panel.
The process covered a long period. This is in accordance with the expectations mentioned in the norm for sensory analysis where it is stated: “Drawing up a sensory profile of a product is a complex procedure and the user of this International Standard needs to know that although this method gives satisfactory results, it requires a large investment in preparation time, calculation and number of training sessions.” This study started with a group of both experienced and inexperienced people. This certainly lengthened the process more than if the panel contained only experienced people. However, the required minimum of six experienced people was not available at the time. By now the process has proven to have been an excellent training to gain evaluation skills.

Has the aim of standardization been reached? The norm states: “Training is satisfactory if each assessor repeats him/herself appropriately (i.e. if the standard deviation is slight for repetitions with the same samples).” And further “It may be that assessors have different perceptions for certain descriptors; in this case, it is important to check that the same assessor evaluates the product in the same way from one session to the next.” The norm does not quantify ‘appropriate repetition’. Has the aim of validation been reached? This study evaluated reliability and validity of the panel according to the authors’ best knowledge of statistical requirements. Based on the described statistical evaluation the conclusion is that the panel is a reliable and appropriate instrument for the visual evaluation of carrot pictures. According to the validity, the conclusion is that the panel is a valid instrument for discriminating pictures of different types of carrots, with increasing power if separate criteria, or clusters of criteria are analysed. The microwave test showed that the panel is able, with a minimum of extra training, to adopt some new criteria and to evaluate new types of pictures.

Concerning the set of criteria used, the PCA’s show four sub-domains. When analysing the discriminating power in these sub-domains, even better results are obtained. The internal consistency is increased in the sub-domains. The described process is unique to the topic biocrystallizations. A solid methodological basis has been laid for the visual evaluation of biocrystallizations and a language for communication has been formed among the scientific community, connected to this method. The present level reached from single morphological features (level 1) towards qualitative descriptive judgements of a higher order, describing gestures in the whole picture (level 3). This is the necessary solid basis to develop the method further towards a judgements on the ‘Gestalt-level’ (level 4), from where the development can be made towards qualitative interpretative judgements of the highest order, namely about food-quality, e.g. authenticity of organic products (level 5). It must be realized that at this interpretative level, a connection needs to be formed between the visual evaluation and the conceptual level, and that on this conceptual level also validation is needed. Next to this, a library of reference material will be needed. It is promising that on the conceptual theme already extensive work has been performed that is useful in connection with the biocrystallization method (Bloksma *et al*., 2001, 2004; Northolt *et al.,* 2004, Huber *et al.,* 2006).

ACKNOWLEDGEMENTS

The authors thank Software AG-Stiftung, EOSTA, Triodos Foundation, Bouwkamp Stichting and Louis Bolk Institute for their financial support to make this project possible. S. Kretschmer is thanked for his great support in finding and modifying the sensory norms, during the work for his master thesis.

References

Amons, F., van der Bie, G. & Huber, M. (1999). Grondslagenonderzoek Proefpersoon-Bloedkristallisatie, kwalitatieve oordelen in de beelddiagnostiek gekwantificeerd. *Eindrapport.* Louis Bolk Institute, Department of Healthcare & Nutrition; Driebergen, The Netherlands.

Andersen, J.-O. (2001). *Development and Application of the Biocrystallization Method*. Report No. 1, Biodynamic Research Association Denmark; Galten, Denmark.

Andersen, J.-O., Henriksen, C., Laursen, J. & Nielsen A. (1999). Computerised image analysis of biocrystallograms originated from agricultural products. *Computers and Electronics in Agriculture*, **22**, *29–48.*

Andersen, J.-O., Huber, M., Doesburg, P., Busscher, N., Mergardt, G., Knijpenga, H. & Waldburger, B. (2005). *Effects of Microwave Heating on the Picture-Developing Properties of Foods, as Examined by Means of a Microwave Degradation Test for Biocrystallizations*. Biodynamic Research Association Denmark; Galten, Denmark.

Balzer-Graf, U. (1991). Steigbild und Kupferchloridkristallisation – Spiegel der Vitalität von Lebensmitteln, In *Lebensmittelqualität, Ganzheitliche Methoden und Konzepte* (A. Meier-Ploeger & H. Vogtmann, eds.), pp. 163–210. Verlag C.F. Müller; Karlsruhe, Germany.

Balzer-Graf, U. (1994). Die Qualität ökologisch erzeugter Produkte. In *Ökologischer Landbau – Perspektiven für die Zukunft* (J. Mayer, O. Faul, M. Ries, A. Gerber & A. Kärcher, eds.). *SÖL*, 261–290.

Balzer-Graf, U. (1996). Vitalqualität von Weizen aus unterschiedlichem Anbau. *Beiträge der biologisch-dynamischen Landwirtschaft*, **44**, 440–450.

Balzer-Graf, U. (2001). Vitalqualität – Qualitätsforschung mit bildschaffenden Methoden, *Ökologie & Landbau*, **117**, 1/2001.

Bessenich, F. (1951). Beiträge zur Erforschung der Bildekräfte durch empfindliche Kristallisation. Naturwissenschaftliche Sektion am Goetheanum; Dornach, Switzerland.

Bijlsma, S. & Bobeldijk, I. (2006). Large-scale human metabolomics studies: A strategy for data (pre-) processing and validation. *Analytical Chemistry,* **78**, 567–574.

Bloksma, J., Northolt, M. & Huber, M. (2001). *Parameters for Apple Quality and an Outline for a New Quality Concept, Part 1 and 2*. Publication FQH-02. Louis Bolk Institute, Department of Healthcare & Nutrition; Driebergen, The Netherlands. Downloadable from www.louisbolk.nl.

Bloksma, J., Northolt, M., Huber, M., Jansonius, P.-J. & Zanen, M. (2004). *Parameters for Apple Quality and the Development of the Inner Quality Concept 2001–2003.* Publication FQH-03. Louis Bolk Institute, Department of Healthcare & Nutrition; Driebergen, The Netherlands. Downloadable from www.louisbolk.nl.

Engquist, M. (1975). *Physische und lebensbildende Kräfte in der Pflanze – ihre Widerspiegelung im Kupferchlorid – Kristallbild*. Vittorio Klostermann; Frankfurt a.M., Germany.
Engquist, M. (1970). *Gestaltkräfte des Lebendigen*, Verlag Vittorio Klostermann; Frankfurt a.M., Germany.
Granstedt, A. & Kjellenberg, L. (1997). Long-term field experiment in Sweden: effects of organic and inorganic fertilisers on soil fertility and crop quality. In *Agricultural Production and Nutrition, Proceedings of an International Conference* (W. Lockeretz, ed.), pp. 79–90. Tufts University; Boston, U.S.A.
Heus, P. de, Leeden, R. van der & Gazendam, B. (2002). *Toegepaste data-analyse: technieken voor niet-experimenteel onderzoek in de sociale wetenschappen.* Reed Business Information; Maarssen, The Netherlands.
Huber, M.A.S., Bloksma, J., Burgt, G.J. van der & Vijver, L.P.L. van de (2006). Challenges for an organic food quality concept – the Inner Quality Concept, Requirements demonstrated on an experimental concept. *Proceedings Odense Joint Conference*, 2006.
Huber, M., Doesburg, P., Andersen, J-O., Paulsen, M., Busscher, N., Kahl, J., Mergardt, G., Kretschmer, S., Zalecka, A., Meelursarn, A., Baars, E. & Nierop, D. (2007). *Validation of Visual Evaluation of Biocrystallizations, Development of a Profiling instrument for Visual Evaluation by a Panel According to ISO-norms for Sensory Analyses.* Triangle report nr. **5**. Louis Bolk Institute, Department of Healthcare & Nutrition; Driebergen, The Netherlands.
ISO (1985a). Sensory Analysis – Methodology – General guidelines. UDC 543.92. Ref.No. ISO 6658–1985 (E).
ISO (1985b). Sensory Analysis – Methodology – Flavour profile methods. Ref.No. ISO 6564 – 1985 (E). ISO (1992). Sensory Analysis – Vocabulary. Ref.No. ISO 5492 : 1992 (E/F). ISO (1994). Sensory Analysis – Identification and selection of descriptors for establishing a sensory profile by multidimensional approach. Ref.No. ISO 11035: 1994 (E).
Kahl, J. (2007). Entwicklung, in-house Validierung und Anwendung des ganzheitlichen Verfahrens Biokristallisation für die Unterscheidung von Weizen-, Möhren-und Apfelproben aus unterschiedlichem Anbau und Verarbeitungsschritten. Habilitationsschrift, Universität Kassel; Witzenhausen, Germany.
Kahl, J. Busscher, N. Meier-Ploeger, A. (2003). Ganzheitliche Untersuchungsmethoden zur Erfassung und Prüfung der Qualität ökologischer Lebensmittel: Stand der Entwicklung und Validierung, Abschlußbericht Projekt 02OE170, Bundesprogramm Ökolandbau.
Kretschmer, S. (2003). Establishing a scientific method according to principles of sensory analysis for the visual evaluation of crystal pictures derived from copper Chloride Crystallization. Master Thesis, University of Kassel, Witzenhausen, Germany.
Landis, J. & Koch, G. (1977a). The measurement of observer agreement for categorical data, *Biometrics*, **33**, 159–174.
Landis, J. & Koch, G. (1977b). A one-way components of variance model for categorical data, *Biometrics*, **33**, 671–679.
Mäder, P., Pfiffner, L., Niggli, U., Balzer, F., Plochberger, K., Velimirov, A. & Besson, J-M. (1993). Effect of three farming systems (bio-dynamic, bio-organic, conventional) on yield and quality of beetroot (*Beta vulgaris* L. var. e*sculenta* L.) in a seven year crop rotation. *Acta Horticulturae*, **339**, 11–31.
Northolt, M., Burgt, G.J. van der, Buisman, T. & Vanden Bogaerde, A. (2004). *Parameters for Carrot Quality and the Development of the Inner Quality Concept.* Publication FQH – 04, Louis Bolk Institute, Department of Healthcare & Nutrition; Driebergen, The Netherlands.
Petterson, B-D. (1967). Beiträge zur Entwicklung der Kristallisationsmethode mit Kupferchlorid nach Pfeiffer. *Lebendige Erde*, **18**, 15–31.
Pfeiffer, E. (1930). *Kristalle*. Orient–Occident Verlag; Stuttgart, Germany.
Schudel, P., Augstburger, F., Eichenberger, M., Vogtmann, H. & Matile, Ph. (1980a). Kompost- und NPK-Düngung zu spinat im Spiegel der empfindlicher Kristallisation und analytischer Daten, Teil 1. *Lebendige Erde*, **31**, 67–70.
Schudel, P., Augstburger, F., Eichenberger, M., Vogtmann, H. & Matile, Ph. (1980b). Kompost- und NPK-Düngung zu spinat im Spiegel der empfindlicher Kristallisation und analytischer Daten, Teil 2. *Lebendige Erde*, **31**, 101–110.
Selawry, A. & Selawry, O. (1957). *Die Kupferchlorid-Kristallisation in Naturwissenschaft und Medizin.* Gustav Fischer Verlag; Stuttgart, Germany.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations; Uses in assessing rater reliability. *Psychological Bulletin*, **86**, 420–428.
Soil Association (2001). *Organic Farming, Food Quality and Human Health*. Soil Association; Bristol, U.K.
Streiner, D. & Norman, G. (1995). *Health Measurement Scales. A Practical Guide to their Development and Use*. 2nd edn. Oxford University Press; Oxford, U.K.
Von Hahn, F. (1962). *Thesigraphie – Untersuchungsmethode zu biologischen Objekten insbesondere Nahrungsmittel*. Franz Steiner Verlag GmbH; Wiesbaden, Germany.
Weibel, F., Bickel, R., Leuthold, S., Alföldi, T., Niggli, U. & Balzer-Graf, U. (2001). Bioäpfel – besser und gesünder? Eine Vergleichsstudie mit Standard und Alternativmethoden der Qualitätserfassung. *Ökologie und Landbau*, **177**, 25–28.Woese, K., Lange D., Boess, C. & Bögl, K. (1995). Ökologisch und konventionell erzeugte Lebensmittel im Vergleich, eine Literaturstudie, Heft 4 und 5, 758 S.
Worthington, V. (1998). Effect of agricultural methods on nutritional quality. A comparison of organic with conventional crops. *Alternative Therapies*, **4**, 58–69.