DOI: 10.18832/kp201803

Methods of Evaluating of Sensory Assessors – Part 1
Způsoby hodnocení senzorických posuzovatelů – část 1

Pavel ČEJKA, Jana OLŠOVSKÁ, Karel ŠTĚRBA, Martin SLABÝ, Tomáš VRZAL
RIBM Plc, Lípová 15, CZ 120 44 Praha 2, Czech Republic
VÚPS, a. s., Lípová 15, 120 44 Praha 2
e-mail: cejka@beerresearch.cz

Recenzovaný článek / *Reviewed Paper*

Čejka, P., Olšovská, J., Štěrba, K., Slabý, M., Vrzal, T., 2018: Methods of evaluating of sensory assessors – part 1. Kvasny Prum. 64 (1): 14–20

The precission and repeatability of sensory analysis is influenced by a number of physiological, psychological and genetic factors. These can be largely eliminated by appropriate selection of the assessors through their regular evaluation. Therefore, assays have been developed to define the divergence of assessor ratings from the rest of the tasting committee. This makes it possible to recognize the area that needs to be focused on in the training of assessors. The article presents specific examples of calculations to determine the degree of conformity among assessors using the ANOVA method and other statistical methods.

Čejka, P., Olšovská, J., Štěrba, K., Slabý, M., Vrzal, T., 2018: Způsob hodnocení senzorických posuzovatelů – část 1. Kvasny Prum. 64 (1): 14–20

Správnost a opakovatelnost senzorické analýzy ovlivňuje řada fyziologických, psychologických a genetických faktorů. Ty lze z větší míry eliminovat vhodným výběrem posuzovatele jeho pravidelným hodnocením. Proto byly vyvinuty způsoby, pomocí nichž lze definovat odlišnost hodnocení posuzovatele od zbytku senzorické komise. To umožňuje rozpoznat oblast, na kterou je třeba se při výcviku posuzovatelů zaměřit. V článku jsou uvedeny konkrétní ukázky výpočtů na určení míry shody mezi posuzovateli pomocí metody ANOVA a dalších statistických metod.

Keywords: *beer, sensory analysis, sensory assessor, testing*

Klíčová slova: *pivo, senzorická analýza, posuzovatel, testování*

1 INTRODUCTION

Maintaining an unchanging characteristic sensory profile of particular beer brand is a challenging task for breweries. An immensely important role in this process is played by the sensory committee, which should reliably determine whether the beer produced corresponds to a defined sensory profile or whether it deviates from it.

High demands are placed on the members of the evaluation committee. The assessor must be able to correctly identify a large number of sensory-active substances that are present in the beer in a concentration range from ppt to hundreds of ppm, i.e. in the range of 9–10 orders. In addition, other requirements, e.g. of a psychological nature (for example, being subject to collective decision-making in a committee, lack or excess of self-confidence, etc.), are placed on the assessor.

From a metrological point of view, the sensory analytical panel is an actual measuring instrument and the results of each analysis performed depend on its members. However, they are less reproducible both in time and in comparison with analytical instruments, moreover, they are more susceptible to errors. This is due to the fact that sensory stimulus perception is an active and selective process; the assessor perceives only the elements of the complex perception he considers relevant, and the rest is forgotten. Extending and improving sensory evaluation can be achieved through regular training. For better targeting of both individual and general tasting sessions, assessments of sensory tasters can reveal the areas that need to be given more attention in training.

2 FACTORS INFLUENCING THE RESULTS OF SENSORY ANALYSIS (ČSN EN ISO 5492)

During sensory assessment, individual testers may be affected by a number of factors that, when neglecting their influence, may be reflected not only in the assessment provided by a particular person but also, in a worse case, in the overall outcome of the tasting of the entire committee. These factors can be divided into physiological, psychological and genetic factors (ČSN EN ISO 5492, 2009).

2.1 Physiological factors

Adaptation (physiological fatigue) is a decrease or change in sensitivity to given stimuli as a result of the continuous action of these stimuli. In sensory evaluations, this is an unwanted source of variability, especially in terms of quantitative intensity assessment. An example may be the evaluation of distilled water after tasting samples with sucrose (perceived as bitter) or caffeine (perceived as sweet).

1 ÚVOD

Udržení neměnného charakteristického senzorického profilu piva je pro pivovary náročným úkolem. Nezastupitelnou roli v tomto procesu hraje senzorická komise, která by měla spolehlivě určit, zda vyrobené pivo odpovídá definovanému senzorickému profilu nebo zda se od něj nějakým způsobem odchyluje.

Na členy posuzovatelské komise jsou kladeny vysoké nároky. Posuzovatel musí být schopen správně identifikovat velké množství senzoricky aktivních látek, které se v pivu vyskytují v koncentračním rozmezí od ppt až po stovky ppm, tedy v rozsahu 9–10 řádů. Kromě toho jsou na posuzovatele kladeny i další požadavky např. psychologického charakteru (podřízení se kolektivnímu rozhodování v komisi, nedostatek nebo přemíra sebevědomí aj.).

Z metrologického hlediska senzorický analytický panel představuje skutečný měřicí přístroj a výsledky každé prováděné analýzy závisí na jeho členech. Ti jsou však oproti přístrojům analytickým méně reprodukovatelní jak v čase, tak i mezi sebou, a jsou i více náchylní k chybám. To je způsobeno tím, že vnímání senzorického podnětu je aktivní a selektivní proces, tzn. posuzovatel vnímá jen ty prvky z komplexního vjemu, které považuje za relevantní, a zbytek pomíjí. Rozšíření a zkvalitnění senzorického hodnocení lze dosáhnout školením a pravidelným tréninkem. Pro lepší zacílení tréninků jak jednotlivců, tak i celé senzorické komise, slouží hodnocení senzorických posuzovatelů, které může odhalit oblasti, jimž je třeba při výcviku věnovat větší pozornost.

2 FAKTORY OVLIVŇUJÍCÍ VÝSLEDKY SENZORICKÉ ANALÝZY

Během senzorického posuzování mohou být jednotliví posuzovatelé ovlivněni řadou faktorů, které se při zanedbání jejich vlivu mohou projevit na výsledku nejen konkrétního člověka, ale v horším případě i na celkovém výsledku degustace celé komise. Tyto faktory lze dle jejich původu rozdělit na fyziologické, psychologické a genetické (ČSN EN ISO 5492, 2009).

2.1 Fyziologické faktory

Adaptace (fyziologická únava) je pokles nebo změna citlivosti k daným stimulům jako výsledek kontinuálního působení těchto stimulů. V senzorických hodnoceních je to nechtěný zdroj variability, a to hlavně ve smyslu kvantitativního hodnocení intenzity. Příkladem může být hodnocení destilované vody po ochutnávání vzorků se sacharosou (je vnímána jako hořká) nebo kofeinem (je vnímána jako sladká).

Poor physical condition, i.e. illness, cold, gum inflammation or other diseases in the oral cavity, is manifested by reduced sensitivity to most stimuli. The assessor condition is also aggravated by emotional stress (at work or at home).

Feeling hungry/satiety – the assessment should take place at least 2 hours after a hearty meal, so that the assessor does not feel overwhelmed, but cannot be hungry or thirsty. Another important condition is that no more beer should be consumed in intervals between series of samples. Smoking does not necessarily aggravate sensory abilities, but it is recommended that the assessors do not smoke at least 30 to 60 minutes before the start of the assessment.

The gender composition of the committee should be balanced as far as possible; women tend to be more sensitive to aromas and tastes than men.

2.2 Psychological factors

Expectancy (autosuggestion) poses a risk that the information on the sample can result in a prediction of the judgment (a person often looks for what he expects to find). For instance, if the assessors receive information that there is a complaint concerning a sample, they tend to judge the sample more strictly. The expectation error significantly exacerbates the validity of the test, and therefore samples must be anonymized (e.g. by using codes without additional information on sample history).

Custom error (sensory blindness) occurs when the sensory property changes slowly over time. The assessor does not notice this gradual change and tends to evaluate intensity of the stimulus as constant. A custom error can be avoided by inserting blind samples.

Stimulus error arises if an irrelevant stimulus changes the attitude of the assessor. For example, appearance (color, texture, packaging, etc.) can suggest better or worse ratings to the assessor. To avoid this error, you need to submit samples in neutral packages and, e.g., ensure that the sample is lighted by a different color.

Halo effect represents a situation where the assessment is influenced by the overall impression (according to generalizing prejudices). For example, if the first overall impression of a sample is positive, its individual attributes are evaluated positively, or if more attributes are evaluated on the sample, the evaluation of one affects the evaluation of the next one. It is therefore preferable to evaluate individual properties separately in one row.

A very important factor is the order of samples, as it can cause an unwanted effect:

The contrast effect – a neutral sample following a high quality one is rated worse and vice versa.

Group effect (opposite to contrast) – a good sample in a series of poor quality samples can be assessed worse and vice versa.

Centralized effect – evaluators tend to rate samples closer to the average.

Positioning effect – samples taken at the beginning (mainly the first sample) are abnormally preferred or rejected.

Mutual suggestion arises if the assessors are influenced by the behavior of other assessors. Therefore, verbal evaluation and other audio stimuls should be avoided. Ideal is the separation in the boxes so that the assessors do not perceive the expressions in the face of others.

Lack of motivation leads to worse results. The motivated assessor is always more efficient. Important is the awareness that the results are important and that the committee has prestige. Assessors should have feedback on their evaluation. A smaller financial reward is also appropriate, although the evaluation is done in the working time (this is also dependent on attendance).

Dominance, shyness, moodiness are features that damage the assessment within the committee. The head of the panel should monitor the behavior of individual assessors (e.g. by using blind samples) and their individual attitude to adapt to the attitudes of the commission as a whole. Within the panel, all assessors have equal status, regardless of their position in the company.

2.3 Genetic factors

Research published over the past 15 years shows that the genetic diversity in olfaction properties among humans is much greater than previously thought. There are plenty of compounds that people perceive with different quality or different intensity. Thus, a particular assessor can respond to odor in a different way than another one. This also applies to a number of known brewing compounds, such as diacetyl or indole. Diacetyl thresholds range from 30-1000 ppb, some of the people being totally anosmic to diacetyl (they do not even perceive it). Indole, which can get into the beer due to contamination

Špatná fyzická kondice, tj. např. nemoc, nachlazení, zánět dásní nebo jiné onemocnění v ústní dutině, se projevuje sníženou citlivostí na většinu podnětů. Kondici posuzovatele také zhoršuje emoční stres (v práci nebo doma).

Pocit hladu/přesycení – hodnocení by se mělo konat minimálně dvě hodiny po vydatném jídle, aby posuzovatel neměl pocit přesycení, ale nesmí být ani hladový ani žíznivý. Další důležitou podmínkou je, že v přestávkách mezi sériemi vzorků nesmí konzumovat další pivo. Kouření nemusí nutně senzorické schopnosti zhoršovat, ale doporučuje se, aby degustátoři nekouřili alespoň 30 až 60 minut před začátkem degustace.

Genderové složení komise by mělo být pokud možno vyrovnané, ženy bývají na některé vůně a chutě citlivější než muži.

2.2 Psychologické faktory

Expektační chyba (autosugesce) představuje riziko v tom, že informace o vzorku mají za následek předjímání úsudku o něm (člověk často hledá to, co očekává, že najde). Např. dostanou-li posuzovatelé vzorek s informací, že se jedná o reklamaci, mají tendenci posuzovat vzorek přísněji. Expektační chyba výrazně zhoršuje validitu testu, a proto musí být vzorky anonymizovány (např. pomocí kódů, bez dodatečných informací o historii vzorku).

Zvyková chyba (senzorická slepota) vzniká, pokud se pomalu v čase mění nějaká senzorická vlastnost. Posuzovatel tuto pozvolnou změnu nezaznamená a má tendenci měnící se intenzitu stimulu hodnotit pořád stejně. Zvykové chybě se lze vyhnout zařazováním slepých vzorků.

Stimulační chyba vznikne, pokud irelevantní podnět(y) změní postoj posuzovatele. Například vzhled (barva, textura, obal apod.) může sugerovat posuzovateli lepší nebo horší hodnocení. Pro zamezení této chyby je třeba předkládat vzorky v neutrálních obalech, zajistit osvit vzorku jinou barvou apod.

Halo efekt představuje situaci, kdy posuzování je ovlivněno celkovým dojmem (podle zobecňujících předsudků). Pokud je například první celkový dojem o vzorku příznivý, i jeho jednotlivé atributy se hodnotí kladně, nebo pokud se na vzorku hodnotí více atributů, hodnocení jednoho ovlivňuje hodnocení dalšího. Bývá proto výhodnější hodnotit jednotlivé vlastnosti samostatně v jedné řadě.

Velmi důležitým faktorem je pořadí vzorků, neboť může vyvolat nechtěný efekt:

Kontrastní efekt – neutrální vzorek následující po velmi kvalitním je hodnocen hůře a naopak.

Skupinový efekt (opačný ke kontrastnímu) – kvalitní vzorek v řadě nekvalitních vzorků může být hodnocen hůře a naopak.

Centrální efekt – posuzovatelé mají tendenci hodnotit vzorky blíže k průměru.

Poziční efekt – vzorky podané na začátku (hlavně první vzorek) bývají abnormálně preferovány nebo naopak odmítnuty.

Vzájemná sugesce vzniká, pokud jsou posuzovatelé ovlivněni chováním ostatních posuzovatelů. Proto je třeba vyloučit slovní hodnocení a další zvukové podněty. Ideální je separace v boxech, aby nevnímali výrazy ve tváři ostatních.

Nedostatek motivace vede k horším výsledkům, motivovaný posuzovatel je vždy výkonnější. Důležité je vědomí, že výsledky jsou důležité a že komise má prestiž. Posuzovatelé by měli mít zpětnou vazbu na svoje hodnocení. Vhodná je také menší finanční odměna, i když se hodnocení koná v pracovní době (závislá též na účasti).

Dominance, plachost, náladovost jsou vlastnosti, které poškozují hodnocení v rámci komise. Vedoucí panelu by měl monitorovat chování jednotlivých posuzovatelů (např. užíváním slepých vzorků) a individuálním přístupem jejich postoje přizpůsobovat postojům komise jako celku. V rámci panelu mají všichni posuzovatelé bez ohledu na jejich pozici ve firmě rovné postavení.

2.3 Genetické vlivy

Výzkumy publikované v posledních asi 15 letech ukazují, že genetická rozmanitost v olfaktorických vlastnostech je mezi lidmi mnohem větší, než se dříve myslelo. Existuje spousta sloučenin, které lidé vnímají různě kvalitativně nebo různě intenzivně. Tedy konkrétní posuzovatel může reagovat na vůni (zápach) odlišným způsobem než jiný posuzovatel. To platí i pro řadu známých pivovarských sloučenin, jako např. diacetyl nebo indol. Prahové hodnoty diacetylu se pohybují mezi 30 – 1000 ppb, přičemž někteří lidé jsou na diacetyl zcela anosmičtí (vůbec jej nevnímají). Indol, který se do piva může dostat kontaminací koliformními bakteriemi, připomíná určité části populace květinovou (jasmínovou) vůni, zatímco pro zbytek populace páchne fekálním materiálem. Někteří lidé vnímají obě vůně indolu v závislosti na jeho koncentraci (Spinney, 2011).

with coliform bacteria, reminds certain parts of the population of a floral (jasmine) fragrance, while for the rest of the population it smells like fecal material. Some people perceive both scents of indole depending on its concentration (Spinney, 2011).

Other major differences are described in the perception of the bitterness of specific substances. There are many substances in the nature that have a bitter taste, such as caffeine, quinine, epicatechin, some alkaloids, sucrose octaacetate and many others. Each compound exhibits different intensity and often a different quality – harsh, adherent, etc. (Drewnowski, 2001). Since bitterness was an evolutionary signal for a possible poison, it is not surprising that the adaptation to the positively perceived bitter taste evolved gradually. Genetic research has shown that there are differences in the perception of PTC (phenylthiocarbamide) and PROP (6-propyl-2-thiouracil), which are bitter compounds used to test the bitterness sensitivity. Approximately one quarter of the population is totally insensitive to these substances (they do not perceive them as bitter), about 50% are moderately sensitive and 25% very sensitive (they perceive them as very bitter, with a higher proportion of these people being represented among Asians and women) (Duffy et al., 2004; Negri et al., 2012). However, for example, people who are sensitive to PROP may be more susceptible to caffeine or other bitter substances (Drewnowski, 2001).

Bitterness is perceived by sensors based on bitter taste receptors, which means that there are specific structures on the tongue that respond to specific molecules. Many different types of receptors respond to a wide variety of bitter compounds. The fact that genes that control susceptibility to specific (but not all) bitter compounds have been discovered, and that mixtures of compounds can have synergistic effects, supports the idea that there are many types of taste receptors whose properties are genetically conditioned (Walter and Roy, 2006; Guinard et al., 1996).

3 SELECTION, TRAINING AND MONITORING OF SENSORY ASSESSORS

Selection and training of assessors is determined in detail by the standard ČSN EN ISO 8586 (560037) "Sensory Analysis – General Guidelines for Selection, Training and Monitoring of Selected Judges and Professional Sensory Judges". This International Standard specifies criteria for the selection, training and monitoring of the work of selected judges and expert selected assessors (experts). (ČSN EN ISO 8586).

3.1 Selection of assessors

Three tasks are encountered when selecting assessors (ČSN EN ISO 8586, 2015): to determine the number of evaluators for the given test, to obtain suitable candidates and to set criteria for their selection.

The number of evaluators should be large enough for statistical evaluation, and ideally should be not less than ten. On the other hand, too many assessors (more than 15) bring problems of organizational and technical nature.

Assessors can be obtained from "internal sources", i.e. employees, or from "external sources". Both options have their positives and negatives and it needs to be decided based on the requirements of evaluation (sample anonymity, connection with production) as well as organization of tasting (availability of tasters, financial difficulty).

The issues when selecting assessors include their time schedules, preference of evaluated samples, knowledge and presumption for tasting, ability to communicate and describe the evaluated sample, and health condition. The psychological characteristics should include their interest and motivation to evaluate, responsibility and the ability to concentrate and, last but not least, the ability of team work.

3.2 Training of the assessors

To make the results of the tasting committee reliable it is necessary for the individual assessors to be constantly tested in addition to the regular training. Evaluation training should be both qualitative and quantitative. In addition, it is necessary to evaluate how the assessor works within the committee.

Qualitative assessment means the ability of the assessor to determine and, most importantly, to correctly name different tastes and smells. This can be performed by adding various compounds to beer; large brewery committees often participate in international testing using tablets added to beer (e.g. FlavorActiV™ or Aroxa™).

Quantitative assessments are made in the form of concentration ranges, where solutions are prepared to increase the concentration

Další velké popsané rozdíly existují ve vnímání hořkosti konkrétních látek. V přírodě existuje velké množství látek, které vykazují hořkou chuť, např. kofein, chinin, epikatechin, některé alkaloidy, oktaacetát sacharosy a mnoho dalších. Každá sloučenina má různou intenzitu a často i jinou kvalitu, včetně drsné, ulpívající atd. (Drewnowski, 2001). Vzhledem k tomu, že hořkost byla evolučním signálem pro možný jed, není překvapením, že adaptace na hořkou chuť vnímanou pozitivně se vyvíjela postupně. Genetický výzkum ukázal, že existují rozdíly ve vnímání PTC (fenylthiokarbamidu) a PROP (6-propyl-2-thiouracilu), což jsou hořké sloučeniny používané k testování citlivosti na hořkost. Přibližně platí, že asi čtvrtina populace je na tyto látky zcela necitlivá (vůbec je nevnímají jako hořké), asi 50 % je středně citlivých a 25 % velmi citlivých (vnímají je jako silně hořké; vyšší procento těchto lidí je zastoupeno mezi Asiaty a ženami) (Duffy et al., 2004, Negri et al., 2012). Přitom např. lidé necitliví na PROP mohou být citlivější na kofein nebo jiné hořké látky (Drewnowski, 2001).

Hořkost je vnímána senzory založenými na chuťových receptorech hořkosti, což znamená, že na jazyku existují specifické struktury, které reagují na specifické molekuly. Široké škále sloučenin, které jsou hořké, odpovídá řada různých druhů receptorů. Skutečnost, že byly objeveny geny, které řídí citlivost na specifické (ale ne všechny) hořké sloučeniny, a také, že směsi sloučenin mohou mít synergické účinky, podporuje myšlenku, že existuje mnoho typů chuťových receptorů, jejichž vlastnosti jsou geneticky podmíněné (Walter a Roy, 2006; Guinard et al., 1996).

3 VÝBĚR, VÝCVIK A SLEDOVÁNÍ SENZORICKÝCH POSUZOVATELŮ

Výběr a výcvik posuzovatelů podrobně určuje ČSN EN ISO 8586 (560037) „Senzorická analýza – Obecná směrnice pro výběr, výcvik a sledování činnosti vybraných posuzovatelů a odborných senzorických posuzovatelů". Tato mezinárodní norma určuje kritéria pro výběr, postup výcviku a sledování práce vybraných posuzovatelů a odborných vybraných posuzovatelů (expertů). (ČSN EN ISO 8586, 2015).

3.1 Výběr posuzovatelů

Při výběru posuzovatelů (ČSN EN ISO 8586, 2015) je třeba vyřešit tři problémy: určit počet posuzovatelů pro danou zkoušku, získat vhodné kandidáty a stanovit kritéria jejich výběru.

Počet posuzovatelů by měl být dostatečně velký pro statistické vyhodnocení, v ideálním případě by neměl být menší než deset. Naopak při příliš vysokém počtu (více než 15) nastávají problémy zejména organizačního a technického rázu.

Posuzovatele lze získávat jak z „vnitřních zdrojů", tzn. zaměstnanců, nebo ze „zdrojů externích". Obě možnosti mají svá pozitiva i negativa a je třeba se rozhodnout na základě potřeb hodnocení (anonymita vzorků, spojení s výrobou) i organizace degustace (dostupnost degustátorů, finanční náročnost).

Při výběru posuzovatelů by se měly vzít do úvahy jejich časové možnosti, obliba typu hodnocených vzorků, znalosti a předpoklady k degustaci, schopnost komunikovat a popsat hodnocený vzorek, zdravotní kritéria, z psychologických vlastností pak jejich zájem a motivace hodnotit, zodpovědnost a schopnost se koncentrovat a v neposlední řadě i schopnost týmové práce.

3.2 Výcvik posuzovatelů

Aby výsledky senzorické komise byly hodnověrné, je třeba, aby jednotliví posuzovatelé byli kromě pravidelného školení neustále trénováni a prověřováni. Trénink hodnocení by měl být jak kvalitativní, tak i kvantitativní. Dále je třeba hodnotit, jak daný posuzovatel pracuje v rámci komise.

Kvalitativním hodnocením se rozumí schopnost posuzovatele určit, a hlavně správně pojmenovat různé chutě a vůně. To lze provádět přídavky různých sloučenin do piva; komise z velkých pivovarů se často zúčastňují mezinárodního přezkušování formou tablet, přidaných do piva (organizují např. FlavorActiV™ nebo Aroxa™).

Kvantitativní hodnocení se provádí formou koncentračních řad, kdy se připraví roztoky o stoupající koncentraci několika chutí (např. sladké, hořké, kyselé), nebo vůní (např. ethylacetát, isoamylacetát, diacetyl, DMS, geraniol atd.). Rovněž některé firmy (Aroxa™) organizují toto hodnocení.

3.3 Hodnocení posuzovatelů

Vzhledem k tomu, že posuzovatelská komise nemůže principiálně poskytovat zcela homogenní výsledky, neměli by se v ní vyskytovat

of several flavors (e.g. sweet, bitter or sour) or aroma (e.g., ethyl acetate, isoamylacetate, diacetyl, DMS, geraniol, etc.). Also, some companies (AroxaTM) organize this testing.

3.3 Evaluation of assessors

Although the evaluation committee cannot, in principle, provide entirely homogeneous results, it should not contain individuals who systematically deviate in certain criteria. One way to limit the occurrence of this phenomenon is a discussion by the participants on the evaluation of the already tested samples (of course, without the possibility of retrospective changes).

Even if the sensory panel leader is able to provide the training of the assessors and optimal evaluation conditions to the maximum possible extent, deviations of individual assessors caused by physiological (e.g. sensitivity and ability to recognize the perception), or psychological factors (e.g. a different way of using an intensity scale caused, among other things, by the lack of concentration of the assessor during the evaluation) will always occur (Naes, 2010). It is therefore appropriate to perform statistical processing of the results at certain time intervals and derive from it how the individual members of the committee agree or differ in terms of the evaluation.

This year, the translation of ISO 11132 (Sensory analysis – Methodology – Guidelines for monitoring the performance of a quantitative sensory panel) into the Czech language, which deals with this problematics (EN ISO 11132, 2012), will be published. This standard provides detailed guidance on how to judge the panel's work.

The standard addresses the following issues:

- Agreement – ability of different panels or assessors to assign similar scores on a given attribute to samples of the same product;
- Homogeneity – measure of the agreement of responses among individual assessors within a test session, as a panel of assessors in replicate sessions, or for an individual assessor in replicate sessions;
- Assessor bias – tendency of an assessor to give scores which are consistently above or below the true score when that is known or the panel mean when it is not;
- Outlier – an assessment that does not conform to the overall pattern of the data or is extremely different from other assessments of the same or similar products;
- Panel drift – phenomenon where a panel, over time, changes in sensitivity or becomes susceptible to biases and as a consequence changes the location on the scale where an attribute is rated for a constant, reference product;
- Performance – ability of a panel or an assessor to make valid and reliable assessments of stimuli and stimulus attributes;
- Repeatability – agreement in assessments of equivalent product samples under the same test conditions by the same assessor or panel;
- Reproducibility – agreement in assessments of equivalent product samples under different test conditions, with different tasks or by a different assessor or panel, in the short term, in the medium or long term;
- Validation – process of establishing that sensory data correlate with other data on samples of the same product (e.g. laboratory measurements, consumer perception, results from other panels, consumer complaints) or that a panel or assessor is able to meet specified performance criteria;
- Session – occasion on which products are assessed;
- Replicate sessions – sessions in which the assessors, the products, the test conditions, and the task are the same.

Performance data monitoring allows the panel manager to improve panel performance and assessors, identify problems and retraining needs, or identify assessors who are not performing well enough to continue to participate. In the following text, there are methods that have proven to be useful in practice.

Variance analysis ANOVA

The analysis of variance (Naes et al., 2010) will be used if we want to find out whether there is a statistically significant difference between the assessors or if we have a large number of rated attributes and we want to determine which attributes are important in terms of the resolution between samples. If the attribute has no effect on product evaluation, it can be stated that the committee is unable to determine the difference between the samples based on this attribute and it can be excluded from further investigation. In the same way, the assessors can be tested for the evaluation of individual attributes to determine which of the attributes to be monitored should

jedinci, kteří systematicky vybočují v určitých kritériích. Jednou z možností, jak výskyt tohoto jevu omezit, je diskuze účastníků hodnocení o již posouzených vzorcích (samozřejmě bez možnosti výsledky zpětně měnit).

I když se vedoucímu senzorického panelu podaří v maximální možné míře zajistit trénink posuzovatelů a optimální podmínky hodnocení, vždy se budou vyskytovat individuální odchylky degustátorů způsobené jednak fyziologickými faktory (např. citlivost a schopnost poznat daný vjem), nebo psychologickými faktory (např. rozdílný způsob použití intenzitní stupnice způsobený mj. nedostatečnou koncentrací posuzovatele během hodnocení. (Naes, 2010) Proto je vhodné v určitých časových intervalech provést statistické zpracování výsledků a z něj odvodit, jak se jednotliví členové komise shodují nebo odlišují v rámci komise.

Letos vyjde překlad normy ISO 11132 (Senzorická analýza – Metodologie – Všeobecný návod pro monitorování a práce kvantitativního senzorického panelu) v českém jazyce, která se touto problematikou zabývá (EN ISO 11132, 2012). Tato norma dává podrobný návod, jak posuzovat práci panelu.

Norma řeší následující problematiku:

- Shoda (agreement) – schopnost různých panelů nebo posuzovatelů přiřadit obdobné skóre daného atributu vzorkům stejného produktu;
- Homogenita (homogenity) – míra souhlasu odpovědí mezi jednotlivými posuzovateli v rámci sezení panelu, a to v opakovaných zasedáních v rámci celého panelu nebo pro individuálního posuzovatele;
- Vychýlení (bias) posuzovatele – tendence posuzovatele k hodnotám, které jsou konzistentně nad nebo pod průměrem skóre;
- Odlehlá hodnota (outlier) – jednotlivé posouzení, které je značně odlišné od ostatních hodnocení stejných nebo podobných výrobků;
- Posun (drift) panelu – jev, kdy panel v průběhu času mění citlivost nebo se stává náchylný k vychýlení a v důsledku toho změní skóre, kterým je atribut hodnocen pro konstantní referenční produkt;
- Interpretace (performance) – schopnost panelu nebo posuzovatele provést platné a spolehlivé posouzení podnětů pro jednotlivé atributy;
- Opakovatelnost (repeatability) – shoda při hodnocení ekvivalentních vzorků produktu za stejných zkušebních podmínek stejným posuzovatelem nebo panelem;
- Reprodukovatelnost (reproducibility) – shoda při hodnocení ekvivalentních vzorků produktu za různých zkušebních podmínek, s různými posuzovateli nebo panely, a to v krátké době, ve střednědobém nebo dlouhodobém horizontu;
- Validace (validation) – proces zjišťování, že senzorická data jsou v souladu s jinými údaji na vzorcích stejného výrobku (např. laboratorní měření, vnímání spotřebitele, výsledky z jiných panelů, stížnosti spotřebitelů) nebo že panel nebo posuzovatel je schopen splnit stanovená specifikovaná kritéria;
- Sezení (session) – časová příležitost, kdy jsou výrobky hodnoceny. V jednom sezení může jeden nebo více produktů hodnotit jeden nebo více posuzovatelů. Posuzovatelé mohou výrobky hodnotit samostatně (v různém čase) nebo ve stejnou dobu jako součást panelu;
- Opakovaná sezení (replicate sessions) – sezení, v nichž jsou posuzovatelé, produkty, zkušební podmínky a úkoly stejné.

Sledování výkonnostních dat umožňuje vedoucímu panelu zlepšit výkonnost panelu a posuzovatelů, identifikovat problémy a potřeby rekvalifikace, nebo identifikovat posuzovatele, kteří nemají dostatečnou výkonnost, aby mohli pokračovat v účasti. V dalším textu jsou uvedeny metody, které se autorům osvědčily v praxi.

Analýza rozptylu – ANOVA

Analýzu rozptylu (Naes et al., 2010) použijeme v případě, když chceme zjistit, zda je mezi posuzovateli statisticky významný rozdíl, nebo pokud máme velké množství hodnocených atributů (vlastností) a chceme zjistit, které atributy jsou důležité z hlediska rozlišení mezi vzorky. V případě, že atribut nevykazuje žádný efekt na hodnocení produktu, lze prohlásit, že panel není schopen určit rozdíl mezi vzorky na základě tohoto atributu a lze jej z dalšího zkoumání vyloučit. Stejně tak lze otestovat i posuzovatele z hlediska hodnocení jednotlivých atributů a určit, které ze sledovaných atributů je třeba podrobit dalšímu testování. Postup při analýze je obdobný jako u vzorků piv, příklad analýzy rozptylu u hodnocení piv uvedli Čejka a Olšovská (2015) u porovnání piv hodnocených pořadovou zkouškou.

Table 1 Evaluation of the overall impression of eight beers by ten assessors
Tab. 1 Hodnocení celkového dojmu osmi piv deseti posuzovateli

Assessor Hodnotitel	Beer / Pivo							
	P1	P2	P3	P4	P5	P6	P7	P8
K	4	3	2	6	4	4	6	4
L	5	2	4	7	5	5	5	5
M	4	3	3	5	5	3	6	4
N	5	3	2	6	4	5	4	5
O	6	4	3	6	3	5	5	5
P	5	3	4	7	5	4	6	4
R	5	4	3	6	5	4	6	5
S	4	3	3	5	6	4	6	5
T	6	6	4	8	5	5	8	6
V	5	4	5	6	4	4	6	5
Average Průměr	4.9	3.5	3.3	6.2	4.6	4.3	5.8	4.8

be subjected to further testing. The analysis procedure is similar to that of beer samples. Čejka and Olšovská (2015) described an approach of the analysis of variance for beer evaluated by the rank test.

Bias – Comparison of the result with the average

The principle of the method (Anděl, 1985) consists in determining the "bias" error which is found from several sensory evaluations. The averages for individual beers are calculated from the respective sensory parameter and subtracted from the evaluations by individual assessors. These values are plotted in the graph or the average and the standard deviation are calculated from them.

Example 1: 10 assessors (designated as K-V, Table 1) evaluated 8 beers (P1-P8). The aim is to find out how individual assessors assess the overall impression (on scale 1 – 9, 1 – best, 9 – worst) within the committee.

By subtracting the averages for individual beers from individual values we obtain the deviation of a particular assessor from the average evaluation of the committee (for assessor K and beer P1 the value is 4 – 4.9 = – 0.9, which means that the assessor rated beer by 0.9 j. better than the committee average). Then, row averages are and the double standard deviation were calculated (Table 2).

The calculated values in the column "average" show the score by which the average assessor evaluates the overall impression in comparison with the average. The column of double SOD is variance range (on one side or the other) in which the assessor moves with 95% probability.

Table 2 then shows that the agreement of assessor R with the committee is the best; he deviates from the overall impression in absolute value by 0.1 and his variance score is only 0.6. On the other hand, the worst match with the committee average is shown by

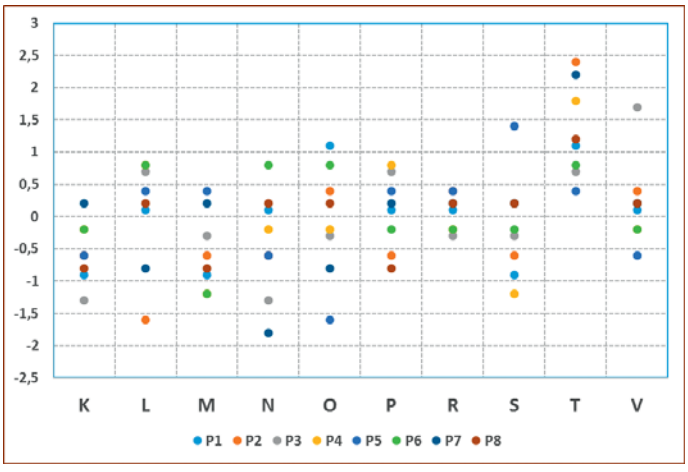

Fig. 1 The score of differences between assessors from the overall impression for 10 evaluators and 8 beers
Obr. 1 Skóre rozdílů hodnocení celkového dojmu pro 10 posuzovatelů a 8 piv

Table 2 Divergence of individual evaluations from average
Tab. 2 Odchylky jednotlivých hodnocení od průměrů

Assessor Hodnotitel	P1	P2	P3	P4	P5	P6	P7	P8	Average Průměr	SOD	2 . SOD
K	-0.9	-0.6	-1.3	-0.2	-0.6	-0.2	0.2	-0.8	-0.6	0.47	0.9
L	0.1	-1.6	0.7	0.8	0.4	0.8	-0.8	0.2	0.1	0.86	1.7
M	-0.9	-0.6	-0.3	-1.2	0.4	-1.2	0.2	-0.8	-0.6	0.60	1.2
N	0.1	-0.6	-1.3	-0.2	-0.6	0.8	-1.8	0.2	-0.4	0.84	1.7
O	1.1	0.4	-0.3	-0.2	-1.6	0.8	-0.8	0.2	-0.1	0.88	1.8
P	0.1	-0.6	0.7	0.8	0.4	-0.2	0.2	-0.8	0.1	0.58	1.2
R	0.1	0.4	-0.3	-0.2	0.4	-0.2	0.2	0.2	0.1	0.28	0.9
S	-0.9	-0.6	-0.3	-1.2	1.4	-0.2	0.2	0.2	-0.2	0.80	1.6
T	1.1	2.4	0.7	1.8	0.4	0.8	2.2	1.2	1.3	0.73	1.5
V	0.1	0.4	1.7	-0.2	-0.6	-0.2	0.2	0.2	0.2	0.68	1.4

Vychýlení – porovnání výsledku s průměrem

Princip metody (Anděl, 1985) spočívá ve stanovení vychýlení (bias), které se zjistí z několika degustací. Z příslušného senzorického parametru se vypočítají průměry pro jednotlivá piva a ten se odečte od hodnocení jednotlivých posuzovatelů. Tyto hodnoty se vynesou do grafu, popř. se z nich vypočte průměr a směrodatná odchylka.

Příklad 1: 10 posuzovatelů (označeni K – V, tab. 1) posuzovalo 8 piv (P1 – P8). Cílem je zjistit, jak jednotliví posuzovatelé posuzují celkový dojem (na stupnici 1 – 9; 1 – nejlepší, 9 – nejhorší) v rámci komise.

Odečtením průměrů pro jednotlivá piva od jednotlivých konkrétních hodnot získáme odchylku konkrétního posuzovatele od průměrného hodnocení komise (pro posuzovatele K a pivo P1 vychází hodnota 4 – 4,9 = – 0,9, což znamená, že tento posuzovatel hodnotil pivo o 0,9 j. lépe, než je průměr komise). Dále se vypočítají řádkové průměry a dvojnásobek výběrové směrodatné odchylky (tab. 2).

Ve sloupci průměr znamenají vypočítané hodnoty, o jaké skóre hodnotí průměrně příslušný posuzovatel celkový dojem oproti průměru. Ve sloupci dvojnásobek SOD je rozptyl (na jednu nebo druhou stranu), ve kterém se posuzovatel s asi 95% pravděpodobností pohybuje.

Z tab. 2 tedy vyplývá, že se nejlépe s komisí shoduje posuzovatel R, který se absolutně liší v celkovém dojmu o 0,1 skóre a rozptyl činí pouze 0,6 skóre. Naopak nejhorší shodu s komisí vykazuje posuzovatel T, který dává horší hodnocení v průměru o 1,3 j. (též obr. 1) a jeho rozptyl je 1,5 j. Přehledně je situace znázorněná na obr. 1.

Hypoteticky mohou nastat čtyři případy, jak výsledky posuzovat (obr. 2):

1) Posuzovatel A: nejlepší soulad s komisí; vykazuje dobrou shodu s průměrem a nízký rozptyl.
2) Posuzovatel B: má tendenci k systematické odchylce, hodnotí daný parametr stále vyšším (nebo nižším skóre), přičemž rozptyl hodnot je nízký.
3) Posuzovatel C: má tendenci poskytovat výsledky s vyšším rozdílem od vztažné hodnoty (rozptyl je vysoký).
4) Posuzovatel D: nejhorší možný případ, je kombinací případů B a C.

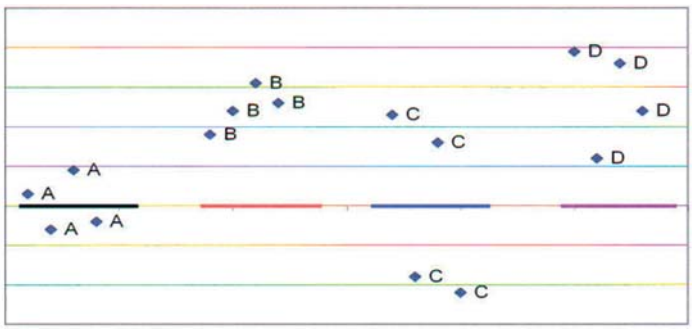

Fig. 2 Ways of evaluating the results of the method of comparison with the average
Obr. 2 Způsoby hodnocení výsledků metody porovnání s průměrem

assessor T, who gives a worse evaluation on average of 1.3 units. (also Fig. 1) and his variance score is 1.5 units. The situation is clearly shown in Fig. 1.

Hypothetically, four cases can occur, how the results can be assessed (Fig. 2):

1) Assessor A: he has the best compliance with the committee; he shows good agreement with average and low variance.
2) Assessor B: he tends to systematically deviate, he evaluates given parameter all the time with a higher (or lower) score, with a low variance of values.
3) Assessor C: he tends to provide results with a higher difference than the reference value, his variance is high).
4) Assessor D: he represents the worst case, it is a combination of cases B and C.

Agreement – The correlation method

Another possibility how to follow the evaluation of the individual assessors is to plot their evaluation data into a chart against the average rating of the whole panel. The values from the previous example are shown in Fig. 3. The values of tasters R (triangle) and T (square) are highlighted. It is clear from the graph that assessor T evaluated all the samples worse than the average mark of the whole committee.

However, when using these methods, it is necessary to have a sufficient difference in the evaluation of the samples. For example, if the average score of all samples ranged between 4 and 5, it is not possible to reliably determine the error of the individual assessors due to the small difference in the values.

Agreement – Box plot

Box plots also allow us to distinguish differences between the assessors. Their advantage is that they represent the variance of individual assessors and are already predefined in the latest versions of MS Excel, so they are readily available. The box plot for example 1 is shown in Fig. 4.

The method of variation coefficient (relative standard deviation)

This method (Anděl, 1985) is used to determine the repeatability of evaluation of individual assessors. Procedure: in several evaluations, the same beer is assigned to the assessor anonymously and for each parameter the variation coefficients are calculated as SOD/average * 100 (%).

Example: Assessor X gave marks for bitterness from 6 evaluations: 2; 3; 3; 3.5; 1.5; 2.5, while assessor Y gave 2.5; 2; 2.5; 1,5; 2; 2.5. Average for X: 2.58; average for Y: 2.17; SOD X: 0.74; SOD Y: 0.41

RSD X = (0.74 / 2.58) .100 = 28%
RSD Y = (0.41 / 2.17) .100 = 19%

Result: assessor Y has a better match in bitterness rating than assessor X.

Shoda – metoda korelace

Další možností, jak sledovat hodnocení jednotlivých degustátorů, je vynést do grafu jejich hodnocení vůči průměrnému hodnocení celého panelu. Hodnoty z předchozího příkladu jsou znázorněny na obr. 3. Zvýrazněné jsou hodnoty degustátora R (trojúhelník) a T (čtverec). Z grafu je dobře patrné, že degustátor T hodnotil všechny vzorky hůře, než byla průměrná známka celé komise.

Při použití těchto metod je však třeba, aby byl dostatečný rozdíl v hodnocení vzorků. Pokud by se např. průměrné hodnocení všech vzorků pohybovalo mezi 4 a 5, nelze díky malému rozdílu hodnot spolehlivě určit chybu jednotlivých posuzovatelů.

Shoda – krabicové grafy

Posuzovat rozdíly mezi posuzovateli umožňují také krabicové grafy. Jejich výhodou je, že znázorňují rozptyl hodnocení jednotlivých degustátorů a jsou již předdefinovány v posledních verzích programu MS Excel, tudíž jsou snadno dostupné. Krabicový graf pro příklad 1 je uveden na obr. 4.

Opakovatelnost – metoda variačního koeficientu (relativní směrodatné odchylky)

Tato metoda (Anděl, 1985) se používá pro zjištění opakovatelnosti hodnocení jednotlivých degustátorů.

Postup: v několika degustacích se zadá posuzovateli anonymně stejné pivo a pro jednotlivé parametry se vypočítají variační koeficienty jako podíl SOD/průměr * 100 (%).

Příklad: Posuzovatel X dal známky pro intenzitu hořkosti z 6 degustací: 2; 3; 3; 3,5; 1,5; 2,5; a posuzovatel Y 2,5; 2; 2,5; 1,5; 2; 2,5.

Průměr X: 2,58; průměr Y: 2,17; SOD X: 0,74; SOD Y: 0,41

RSD X = (0,74/2,58).100 = 28 %
RSD Y = (0,41/2,17).100 = 19 %

Výsledek: posuzovatel Y vykazuje lepší shodu v hodnocení hořkosti než posuzovatel X.

3.4 Opatření vyplývající z hodnocení posuzovatelů

Z výsledků jednotlivých testů lze zjistit, zda má posuzovatel problém s některým ze sledovaných parametrů a následně se v rámci zpětné vazby pro konkrétního degustátora nebo při tréninku komise na toto hodnocení zaměřit.

V případě, že se v komisi vyskytuje degustátor, který se i po pečlivém tréninku dlouhodobě svými výsledky liší od zbytku komise, je třeba takového posuzovatele buď vyměnit, nebo v krajním případě zařadit metodu, která jeho vliv na celkové hodnocení výrazně sníží. Jedním z nejběžnějších způsobů je metoda ořezaného průměru. Nejjednodušší použití této metody spočívá v seřazení hodnot podle velikosti a škrtnutí nejvyšší a nejnižší známky, čímž se eliminují hodnoty nejvíce vzdálené od průměrného hodnocení komise jako celku. Je třeba ale mít vždy na paměti, že pro tuto metodu je třeba mít dostatečný počet hodnocení (ideálně alespoň deset). Pokud se degustace účastní např. pouze tři posuzovatelé, použití metody ořezaného průměru již tak dost nízkou spolehlivost výsledku úplně eliminuje.

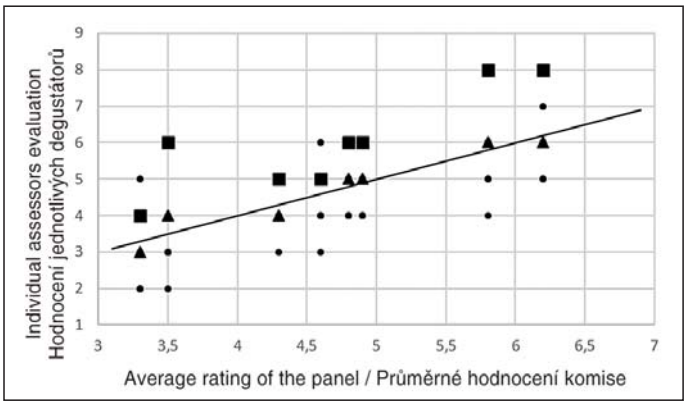

Fig. 3 Correlation of the assessors with the overall impression of the committee (highlighted assessors: R – triangle, T – square)
Obr. 3 Korelace degustátorů vůči celkovému hodnocení komise (zvýraznění degustátoři R – trojúhelník a T – čtverec)

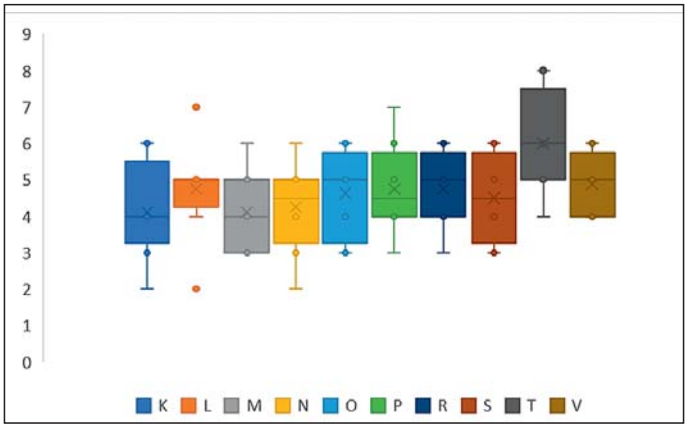

Fig. 4 A box plot of the overall impression rating
Obr. 4 Krabicový graf hodnocení celkového dojmu

3.4 Procedure resulting from the evaluation of the assessors

The results of the individual tests can be used to determine whether the assessor has a problem with some of the monitored parameters and, subsequently, to focus on this problem in the context of the feedback for a particular assessor or the training of the committee.

In the event that an assessor is present in the committee who, differs from the rest of the committee after a careful training session, it is necessary either to replace that assessor or, in the extreme case, to include a method that will significantly reduce its impact on the overall score. One of the most common ways is a trimmed average method. The simplest use of this method is to sort the values by the size and eliminate the highest and lowest values. This eliminates the values that are most distant from the average rating of the committee as a whole. However, it should always be remembered that the use of this method requires a sufficient number of ratings (ideally at least ten). If, for example, only three assessors participate in the tasting, the use of the trimmed average method completely eliminates the already low reliability of the result.

4 CONCLUSIONS

Sensory analysis is an integral part of the evaluation of products in the brewing industry. Its objectivity is ensured by selecting appropriate sensory assessors and their continuous training („calibration"), but also by analyzing their evaluation and taking into account the results of the analysis during the next training.

ACKNOWLEDGEMENTS

This work was developed with the support of CR Ministry of Education Youth and Sport LO1312, project "Sensory Research Centre in Prague and Brewhouse for research and development – sustainability and development (2014–2019, MSM/LO)".

REFERENCES / LITERATURA

Anděl, J., 1985: Matematická statistika. SNTL – Nakladatelství technické literatury, Praha.

Aroxa™: www.aroxa.com/beer. Accessed 17.11.2017

Čejka, P., Olšovská, J., 2015: Využití senzorické analýzy piva v marketingu. Kvasny Prum., 61: 38–45.

ČSN EN ISO 8586, 2015: Senzorická analýza – Obecná směrnice pro výběr, výcvik a sledování činnosti vybraných posuzovatelů a odborných senzorických posuzovatelů. Česká technická norma.

ČSN EN ISO 5492, 2009: Senzorická analýza – Slovník. Česká technická norma.

Drewnowski, A., 2001: The Science and Complexity of Bitter Taste. Nutrition Reviews, 59(6): 163–169.

Duffy, V. B., Davidson, A. C., Kidd, J. R., Kidd, K. K., Speed, W. C., Pakstis, A. J., Reed, D. R., Snyder, D. J., and Bartoshuk, L. M.: Bitter Receptor Gene (TAS2R38), 6-n-Propylthiouracil (PROP) Bitterness and Alcohol Intake. Alcohol Clin. Exp. Res. 2004 Nov; 28(11): 1629–1637.

EN ISO 11132, 2012: Sensory analysis – Methodology – Guidelines for monitoring the performance of a quantitative sensory panel

4 ZÁVĚR

Senzorická analýza je nedílnou součástí hodnocení výrobků v pivovarském průmyslu. Její objektivita je zajišťována jak výběrem vhodných senzorických posuzovatelů a jejich soustavným tréninkem („kalibrací"), ale i analýzou jejich hodnocení a zohledněním výstupů analýzy při dalším tréninku.

PODĚKOVÁNÍ

Tato práce byla vypracována za podpory projektu LO1312 „Výzkumné senzorické centrum v Praze a Výzkumná a vývojová varna – udržitelnost a rozvoj", MŠMT, Česká Republika.

FlavorActiV™: www.flavoractiv.com. Accessed: 25.08.2017

Guinard, J. X., Zoumas-Morse, Ch., Dietz, J., Goldberg, S., Holz, M., Heck, E., Amoros, A., 1996: Does Consumption of Beer, Alcohol, and Bitter Substances Affect Bitterness Perception?, Physiology and Behavior, 59: 625–631.

Naes, T., Brockhoff, P. B., Tomic, O., 2010: Statistics for sensory and consumer science. John Wiley and sons Ltd., Antony Rowe Ltd., Chippenhame, Wiltshire, UK. ISBN 978-0-470-51821-2.

Negri, R., Di Feola, M., Di Domenico, S., Scala, M. G., Artesi, G., Valente, S., Smarrazzo, A., Turco, F., Morini, G., Greco, L., 2012: Taste Perception and Food Choices. J. of P. Gastroenterology and Nutrition, 54: 624–629.

Spinney, L.: You Smell Flowers, I Smell Stale Urine. Scientific America, February 1, 2011

Walters, E., Roy, G., 1996: Taste Interactions of Sweet and Bitter Compounds. Flavor-Food Interactions, Chapter 12: 130–142.

Translated: Karel Sigler
Manuscript received: 16/11/2017
Accepted for publication: 15/12/2017