Journal of Water & Health

Journal of Water and Health Vol 20 No 6, 1005 doi: 10.2166/wh.2022.060

Interlaboratory comparison of the intensity of drinking water odor and taste by two-way ordinal analysis of variation without replication

Tamar Gadrich [a], Ilya Kuselman [b,*], Francesca R. Pennecchi [c], D. Brynn Hibbert [d], Anastasia A. Semenova [e], Pui Sze Cheow [f] and Vladimir N. Naidenko [g]

[a] Department of Industrial Engineering and Management, ORT Braude College, P.O. Box 78, 51 Snunit St., Karmiel 2161002, Israel
[b] Independent Consultant on Metrology, 4/6 Yarehim St., Modiin 7176419, Israel
[c] Istituto Nazionale di Ricerca Metrologica (INRIM), Strada delle Cacce 91, Turin 10135, Italy
[d] School of Chemistry, UNSW Sydney, Sydney, NSW 2052, Australia
[e] V.M. Gorbatov Federal Research Center for Food Systems, 26 Talalikhina St., Moscow 109316, Russia
[f] Health Science Authority, 1 Science Park Road, #01-05/06, The Capricorn, Singapore Science Park II, 117528 Singapore
[g] Ural Research Institute for Metrology – Affiliated Branch of D.I. Mendeleev Institute for Metrology, Krasnoarmeyskaya 4, Ekaterinburg 620075, Russia
*Corresponding author. E-mail: ilya.kuselman@bezeqint.net

TG, 0000-0001-6707-7510; IK, 0000-0002-5813-9051; FRP, 0000-0003-1328-3858; DBH, 0000-0001-9210-2941; AAS, 0000-0002-4372-6448; PSC, 0000-0002-3319-007X; VNN, 0000-0002-9592-2766

ABSTRACT

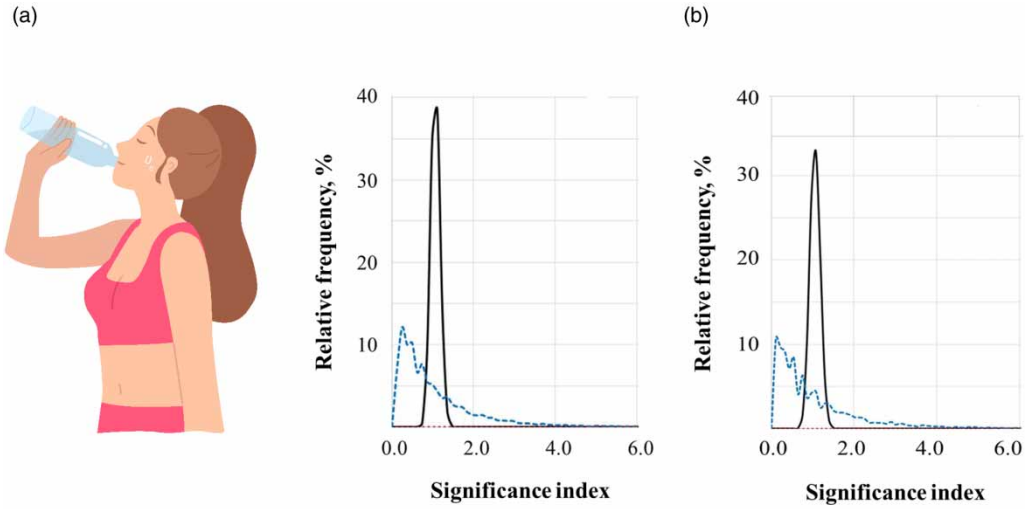
A case study of ordinal data from human organoleptic examination (sensory analysis) of drinking water obtained in an interlaboratory comparison of 49 ecological laboratories is described. The recently developed two-way ordinal analysis of variation (ORDANOVA) is applied for the first time for the treatment of responses on the intensity of chlorine and sulfurous odor of water at 20 and 60 °C, which is classified into the six categories from 'imperceptible' to 'very strong'. The one-way ORDANOVA is used for the analysis of the 'salty taste' intensity of the water. A decomposition of the total variation of the ordinal data and simulation of the multinomial distribution of the data-relative frequencies in different categories allowed the determination of the statistical significance of the difference between laboratories in classifying chlorine or sulfurous odor intensity by categories, while the effect of temperature was not significant. No statistical difference was found between laboratories on salty taste intensity. The capabilities of experts to identify different categories of the intensity of the odor and taste are also evaluated. A comparison of the results obtained with ORDANOVA and ANOVA showed that ORDANOVA is a more useful and reliable tool for understanding categorical data such as the intensity of drinking water odor and taste.

Key words: drinking water, interlaboratory comparison, odor intensity, ORDANOVA, ordinal data, taste intensity

HIGHLIGHTS

- An interlaboratory comparison of human responses to water odor and taste is reported.
- The two-way ORDANOVA is applied for the study of responses to water odor intensity.
- Interlaboratory data on water taste intensity are studied with the one-way ORDANOVA.
- The significance of some factors' influence on a laboratory result is evaluated.
- The applicability of ORDANOVA and ANOVA for ordinal data analysis is discussed.

GRAPHICAL ABSTRACT



Examination of chlorine (a) and sulfurous (b) odor intensity of a drinking water. Empirical distribution functions for testing influence of a laboratory effect (solid line) and a temperature effect (dashed line) on a human response.

INTRODUCTION

The examination of drinking water intensity of odor and taste is important as a foreign odor or taste may indicate water pollution or insufficient purification, besides influencing the aesthetic feelings of a consumer, even if the water is harmless (Burlingame *et al.* 2017). It is equally important to compare the examination responses of experts from different laboratories, i.e., to evaluate how similar or different they are.

Interlaboratory comparisons of quantitative property values (such as a component concentration or content in a substance or material) are widely used for the quality assurance of chemical analytical laboratories including proficiency testing (ISO 17043 2010), validation of analytical methods (Magnusson & Ornemark 2014), and for other purposes (ISO 17025 2017). The standardized statistical techniques for corresponding experiment design and treatment of quantitative continuous data are mostly based on the analysis of variance (ANOVA). At the same time, statistical techniques for interlaboratory comparisons of qualitative (nominal) and semi-quantitative (ordinal) properties of a substance, material, or object are less studied and not harmonized (Tiikkainen *et al.* 2022).

A nominal property of a substance, material, or object is described by a word or alphanumerical code identifying the instance of the property, where the property has existence but no magnitude, e.g., water odor or taste according to human sense (da Silva & Ellison 2021; Hibbert *et al.* 2021). Nominal properties are coded by exhaustive and disjointed classes or categories with no natural ordering. Therefore, nominal data are related to categorical data (Agresti 2012), for which the only legitimate operations are equality or nonequality.

An ordinal property is described by data for which a total ordering relation can be established, according to magnitude, with other quantities of the same kind but for which no algebraic operations exist among those quantities (Hibbert *et al.* 2021). These data are also categorical. Their legitimate operations can be 'equal/unequal' and 'greater/less than'. Examples of such relations are the intensity of an odor and taste. Note that in contrast to kinds/categories of odor (aromatic, marsh, woody, etc.) and taste (bitter, salty, sweet, etc.) having no order, their intensity levels/categories (weak, noticeable, strong, etc.) are ordered.

As the addition of categorical data is not a legitimate operation by definition, whereas one of the ANOVA assumptions is that the factor effects are additive (Scheffé 1999), statistical techniques based on ANOVA cannot be applied directly to nominal and ordinal data.

Possibly, the first statistical technique for the treatment of nominal data, similar to the one-way ANOVA for quantitative continuous data, was developed in the last century (Light & Margolin 1971) and was called 'categorical ANOVA' or CATANOVA. The idea of this technique was to calculate the number of examination responses for the property related to the same category and then to analyze their relative frequency as a fraction of the total number of examination responses for all categories.

Statistical analysis of data obtained in an interlaboratory comparison for a binary nominal and ordinal property (with the number of categories $K = 2$) using the one-way ordinal analysis of variation (ORDANOVA) was proposed by Bashkansky *et al.* (2012), Gadrich & Bashkansky (2012), and Gadrich *et al.* (2013).

The two-way CATANOVA for two variables and $K \geq 2$ categories was developed recently and demonstrated with an interlaboratory comparison of nominal data of macroscopic examinations of weld imperfections (Gadrich *et al.* 2020). The two-way ORDANOVA (Gadrich & Marmor 2021), which was developed simultaneously, is applied in the present paper for the first time to an interlaboratory comparison of ordinal data from a human organoleptic examination of the intensity of odor and taste of drinking water – a kind of sensory data (Hibbert 2020).

Note that odor and taste are important properties of water quality, increasingly attracting the attention of researchers (Lin *et al.* 2019). A search within the *Journal of Water and Health* shows 36 published articles on the topic. The special issue 'Water taste and odor: challenges, gaps, and solutions' was recently announced (Kaloudis *et al.* 2021) in the Elsevier journal 'Chemical Engineering Journal Advances'. The methodology of examination of water odor and taste is a subject of standardization (ISO 20612 2007; GOST 57164 2016; Baird *et al.* 2018). However, we can find no paper or report on an interlaboratory comparison of sensory responses to the intensity of drinking water odor and taste.

The case study analyzed in the present paper was organized in 2020 by the Ural Research Institute for Metrology (UNIIM) – Affiliated Branch of D.I. Mendeleev Institute for Metrology, Russia. Forty-nine Russian ecological laboratories participated in the comparison. Examinations of the intensity of odor and taste of drinking water test items were performed according to the standard (GOST 57164 2016), setting $K = 6$ intensity categories for both the water properties: (a) imperceptible, (b) very weak, (c) weak – does not cause a disapproving response about the water, (d) noticeable – causes a disapproving response, (e) distinct – a tester wishes not to drink, and (f) very strong – the water is not potable. To each category, the standard assigns the respective numeric value (score): 0, 1, 2, 3, 4, and 5. The technical specifications (ISO 20612 2007) for interlaboratory comparisons in the field of water quality, as well as the general guidelines for sensory analysis (ISO 8586 2014), recommend the use of these scores as quantitative responses applying ANOVA or another known statistical technique.

The aim of the present paper is to provide a case study of the intensity of drinking water odor and taste using the ORDANOVA implementation for interlaboratory comparisons of ordinal properties.

THEORY – PRINCIPLES OF ORDANOVA

Layouts

A random phenomenon Y, e.g., an expert response, showing instances on an ordinal scale with K ordered categories/classes/levels is characterized by a probability vector $\boldsymbol{p} = (p_1, p_2, \ldots, p_K)$, where p_k at $k = 1, 2, \ldots, K$ denotes the theoretical probability of responses related to the kth category $\left(\sum_{k=1}^{K} p_k = 1\right)$. Let F_k denote the cumulative theoretical probability up to the kth category, $F_k = \sum_{l=1}^{k} p_l$ and $F_K = 1$. The probability P of receiving a set of responses $(n_1, n_2, \ldots, n_K)$, where n_k $(k = 1, 2, \ldots, K)$ denotes the number of responses related to the kth category, and $\sum_{k=1}^{K} n_k = N$ is calculated based on the multinomial distribution of parameters $(N, \boldsymbol{p})$ as the probability mass function (NIST/SEMATECH 2021):

$$P(n_1, n_2, \ldots, n_K) = \frac{N!}{n_1! \cdot n_2! \cdot \ldots \cdot n_K!} p_1^{n_1} \cdot p_2^{n_2} \cdot \ldots \cdot p_K^{n_K} \tag{1}$$

where $\sum_{k=1}^{K} p_k = 1$.

In the general context, the phenomenon of variability (i.e., variability in the responses of the ordinal variable Y) is taken as explained by two independent factors (random variables) and their possible interaction. In the present work, a particular case is studied where no interaction between the two factors can be analyzed, since only one expert response at the specified levels

of the factors (e.g., at each cell in the cross-balanced design) is examined from each laboratory as required in laboratory proficiency testing (ISO 17043 2010). The first factor, the random variable $X1$, has I levels (for example, I laboratories are discussed), and the second factor, the random variable $X2$, has J levels (e.g., responses are to be received at J different temperatures). There are N responses in total, each of them falling into one of the K categories of the responses of variable Y. On the other hand, each of the N responses falls into one of the I levels of the first factor $X1$ and into one of the J levels of the second factor $X2$. In a cross-balanced design, it is assumed that each of the $I \times J$ cells contains n replicated responses distributed between the K categories. One expert response from a laboratory means $n = 1$, i.e., a cross-balanced design without replication. The frequency n_{ijk} denotes the number of responses in cell (i, j) classified to the kth category $\left(\sum_{k=1}^{K} n_{ijk} = n\right)$, and in total, there are $I \cdot J \cdot n = N$ responses. When $n = 1$, the total number of responses is $I{\cdot}J = N$.

Treating N responses as a statistical sample, and n_{ijk} as a random variable, then, $\hat{p}_{ijk} = n_{ijk}/n$ and $\hat{F}_{ijk} = \sum_{l=1}^{k} \hat{p}_{ijl}$ denote the sample relative frequency of responses belonging to the kth category and the sample cumulative relative frequency of responses up to the kth category in cell (i, j), respectively. The sample total cumulative relative frequency of all responses belonging to the kth category is denoted by

$$\hat{F}_{..k} = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \hat{F}_{ijk} \quad (k = 1, 2, \ldots, K), \tag{2}$$

where $\hat{F}_{i.k} = 1/J \sum_{j=1}^{J} \hat{F}_{ijk}$ $(i = 1, 2, \ldots, I; k = 1, 2, \ldots, K)$; and $\hat{F}_{.jk} = 1/I \sum_{i=1}^{I} \hat{F}_{ijk}$ $(j = 1, 2, \ldots, J; k = 1, 2, \ldots, K)$ denote the sample total cumulative relative frequency of responses up to the kth category at level i of factor $X1$ and at level j of factor $X2$, respectively.

Decomposition of total variation

The total sample variation of the response variable Y, normalized to the [0, 1] interval, is defined in the two-way ORDANOVA model (Gadrich & Marmor 2021) as

$$\hat{V}_{\mathrm{T}} = \frac{1}{(K-1)/4} \sum_{k=1}^{K-1} \hat{F}_{..k}(1 - \hat{F}_{..k}). \tag{3}$$

In the model without replication, the total sample variation $\hat{V}_{\mathrm{T}}$ is partitioned into the between (inter) covariation component $\hat{C}_{\mathrm{B}}$ and the within (intra) residual variation $\hat{V}_{\mathrm{W}}$. For example, in an interlaboratory comparison, the variation $\hat{C}_{\mathrm{B}}$ characterizes the between-laboratory variation of the responses, while the variation $\hat{V}_{\mathrm{W}}$ is the within-laboratory variation. That is

$$\hat{V}_{\mathrm{T}} = \hat{C}_{\mathrm{B}} + \hat{V}_{\mathrm{W}}, \tag{4}$$

where

$$\hat{C}_{\mathrm{B}} = \frac{1}{(K-1)/4} \sum_{k=1}^{K-1} \left[\frac{1}{I} \sum_{i=1}^{I} (\hat{F}_{i.k} - \hat{F}_{..k})^2 + \frac{1}{J} \sum_{j=1}^{J} (\hat{F}_{.jk} - \hat{F}_{..k})^2 \right] \tag{5}$$

and

$$\hat{V}_{\mathrm{W}} = \frac{1}{(K-1)/4} \sum_{k=1}^{K-1} \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} (\hat{F}_{i.k} + \hat{F}_{.jk} - \hat{F}_{..k})(1 - [\hat{F}_{i.k} + \hat{F}_{.jk} - \hat{F}_{..k}]). \tag{6}$$

The individual effects of factors $X1$ and $X2$ can be evaluated using the following decomposition of the variation $\hat{C}_{\mathrm{B}}$:

$$\hat{C}_{\mathrm{B}} = \hat{C}_{X1}^{\mathrm{B}} + \hat{C}_{X2}^{\mathrm{B}}, \tag{7}$$

where

$$\hat{C}^{\mathrm{B}}_{X1} = \frac{1}{(K-1)/4}\sum_{k=1}^{K-1}\frac{1}{I}\sum_{i=1}^{I}\left(\hat{F}_{i.k} - \hat{F}_{..k}\right)^2 \quad \text{and} \quad \hat{C}^{\mathrm{B}}_{X2} = \frac{1}{(K-1)/4}\sum_{k=1}^{K-1}\frac{1}{J}\sum_{j=1}^{J}\left(\hat{F}_{.jk} - \hat{F}_{..k}\right)^2. \tag{8}$$

Another $\hat{C}_{\mathrm{B}}$ decomposition, helpful for comparing the capability of the participating laboratories (as a group) to identify different categories, consists of evaluating the following kth parts of $\hat{C}_{\mathrm{B}}$:

$$\hat{C}_{\mathrm{B}}(k) = \frac{1}{I}\sum_{i=1}^{I}\left(\hat{F}_{i.k} - \hat{F}_{..k}\right)^2 + \frac{1}{J}\sum_{j=1}^{J}\left(\hat{F}_{.jk} - \hat{F}_{..k}\right)^2. \tag{9}$$

Larger values of $\hat{C}_{\mathrm{B}}(k)$ indicate a weaker capability to identify category k. Note that the capability, characterizing dispersion of the responses related up to category k, is analogous to the measurement reproducibility (Hibbert *et al.* 2021). When the cumulative relative frequencies achieve 1, the variation by Equation (9) is 0.

The fraction $\hat{R}^2_{\mathrm{B}}$ of the total sample variation $\hat{V}_{\mathrm{T}}$ reflecting the between-laboratory effect on the response Y is defined as

$$\hat{R}^2_{\mathrm{B}} = \frac{\hat{C}_{\mathrm{B}}}{\hat{V}_{\mathrm{T}}} \qquad (0 \le \hat{R}^2_{\mathrm{B}} \le 1). \tag{10}$$

Similar fractions of the total sample variation reflecting effects of the two factors are:

$$\hat{R}^2_{X1} = \frac{\hat{C}^{\mathrm{B}}_{X1}}{\hat{V}_{\mathrm{T}}}; \quad \hat{R}^2_{X2} = \frac{\hat{C}^{\mathrm{B}}_{X2}}{\hat{V}_{\mathrm{T}}}. \tag{11}$$

The calculations of frequencies, relative frequencies, and variation components can be easily performed using a Microsoft Excel spreadsheet.

Criteria for testing hypotheses on the significance of effects

The null hypothesis of homogeneity of the responses states that the probability of classifying the responses as belonging to the kth category does not depend on the levels of the first factor (levels i) nor on those of the second factor (levels j), i.e., $p_{ijk} = p_k$ for all $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$. Under this hypothesis, the following relations are applicable:

$$\frac{E(\hat{V}_{\mathrm{T}})}{df_{\mathrm{T}}} = \frac{E(\hat{V}_{\mathrm{W}})}{df_w} = \frac{E(\hat{C}^{\mathrm{B}}_{X1})}{df_{X1}} = \frac{E(\hat{C}^{\mathrm{B}}_{X2})}{df_{X2}} = \frac{\dfrac{1}{(K-1)/4}\displaystyle\sum_{k=1}^{K-1}F_k(1-F_k)}{N}, \tag{12}$$

where E is the expected value; $df_{X1} = I - 1$, $df_{X2} = J - 1$, $df_W = (I-1)(J-1)$, $df_{\mathrm{T}} = N - 1$ are the degrees of freedom. The numerator of the last term in Equation (12) is equal to the population total ordinal variation corresponding to the probability vector $\boldsymbol{p} = (p_1, p_2, \ldots, p_K)$.

To check the statistical significance of both the factor effects, the following significance indices (test statistics) have been defined:

$$\widehat{SI}_{X1} = \frac{\hat{C}^{\mathrm{B}}_{X1}/df_{X1}}{\hat{V}_{\mathrm{T}}/df_{\mathrm{T}}}; \quad \widehat{SI}_{X2} = \frac{\hat{C}^{\mathrm{B}}_{X2}/df_{X2}}{\hat{V}_{\mathrm{T}}/df_{\mathrm{T}}}. \tag{13}$$

Testing the null hypothesis H_0 on the effect significance requires the knowledge of at least the asymptotical distribution of the index $\widehat{SI}$ for the calculation of the critical values of the indices at a given level of confidence $(1-\alpha)\cdot 100\%$.

A calculator tool for this purpose was proposed for the two-way ORDANOVA in Gadrich & Marmor (2021). The tool calculates from the empirical data the sample vector of relative frequencies $\hat{\boldsymbol{p}} = (\hat{p}_{..1}, \hat{p}_{..2}, \ldots, \hat{p}_{..K})$, as well as the variation

components ($\hat{C}^{B}_{X1}$, $\hat{C}^{B}_{X2}$, $\hat{V}_W$, $\hat{V}_T$), and the empirical significance indices ($\widehat{SI}_{X1}$, $\widehat{SI}_{X2}$). The critical values SI_{crit} for the indices in Equation (13) are recovered through a Monte Carlo simulation based on at least 10,000 trials. At each iteration, the calculator performs n random draws from the multinomial distribution with K categories and the vector of relative frequencies $\hat{\boldsymbol{p}} = (\hat{p}_{..1}, \hat{p}_{..2}, \ldots, \hat{p}_{..K})$. Calculated significance indices are stored at each realization. Finally, for each significance index, an empirical cumulative distribution function (CDF) is constructed and a relative frequency (%) plot of the simulated $\widehat{SI}$ values (empirical distribution of $\widehat{SI}$) is displayed. The critical value SI_{crit} of the significance index $\widehat{SI}$ is determined as the point where $(1-\alpha)\cdot$ 100 % level of confidence of the empirical CDF is achieved. This corresponds to the $\widehat{SI}$ value on the plot of the relative frequency at which $(1-\alpha)\cdot$ 100% of the area under the curve is cumulated. The null hypothesis H_0 is rejected when the significance index $\widehat{SI}$ exceeds the critical value SI_{crit} at the $(1-\alpha)\cdot$ 100 % level of confidence, concluding that a statistically significant effect on the response variable Y is detected.

The calculator developed using Visual Basic for Applications in a Microsoft Excel spreadsheet is freely available on the link (Marmor & Gadrich 2019).

One-way ORDANOVA

The one-way ORDANOVA can be considered as a simplification of the two-way ORDANOVA when the second factor $X2$ has only one level ($J = 1$). The variability of the responses of the ordinal variable Y is hence explained by a laboratory effect only. The laboratory factor $X1$ has I levels, i.e., I laboratories participate in an interlaboratory comparison.

Assume there are N responses in total, each of them falling into one of the K categories of the variable Y. On the other hand, each of the N responses relates to one of the I laboratories. In the balanced design, n responses from each laboratory (n replicates) are distributed among K categories. The frequency n_{ik} denotes the number of responses from the ith laboratory classified as related to the kth category $\left(\sum_{k=1}^{K} n_{ik} = n\right)$. Hence, the total number is $I \cdot n = N$. In case, $n = 1$, the within (intra) laboratory variation cannot be estimated.

Again, treating N responses as a statistical sample, and n_{ik} as a random variable, then $\hat{p}_{ik} = n_{ik}/n$ and $\hat{F}_{ik} = \sum_{l=1}^{k} \hat{p}_{il}$ denote the sample relative frequency of responses belonging to the kth category and the sample cumulative relative frequency up to the kth category, respectively, in ith laboratory. The sample cumulative relative frequency of responses belonging to the kth category is denoted by $\hat{F}_{.k} = 1/I \sum_{i=1}^{I} \hat{F}_{ik}$ $(k = 1, 2, \ldots, K)$.

Decomposition of the total variation and testing the null hypothesis H_0 of significance of the laboratory effect can also be simplified from a two-way to a one-way ORDANOVA.

EXPERIMENTAL

Preparation of test items

Two test items, 1 and 2, were prepared at UNIIM for the examination of the intensity of chlorine and sulfurous odor, respectively. The components of these items were purchased bottled drinking water (from the same producer and batch), 330 cm^3 in a plastic container for each test item, and the initial solutions of the pure reagents in glass vials: 3 cm^3 of sodium hypochlorite, 0.544 g/dm^3, for test item 1 providing chlorine odor, and 3 cm^3 of sodium sulfide, 0.167 g/dm^3, for test item 2 providing sulfurous odor.

The solution of sodium hypochlorite was mixed with the drinking water before use by each participating laboratory to obtain the final concentration of sodium hypochlorite in test item 1 equal to 4.9 mg/dm^3. This concentration of sodium hypochlorite corresponds to intensity level 2 of chlorine odor, interpolated between levels 1 and 3 described in the standard (GOST 57164 2016).

The final concentration of sodium sulfide in test item 2 equal to 1.5 mg/dm^3 was obtained by mixing its initial solution with the drinking water before use by each participating laboratory. This concentration of sodium sulfide corresponds to intensity level 4 of sulfurous odor, interpolated between levels 3 and 5 by GOST 57164 (2016).

Test item 3 for the examination of intensity of 'salty taste' was prepared at UNIIM as 330 cm^3 of sodium chloride solution in the drinking water in a plastic container (0.73 g/dm^3) corresponding to intensity level 2 of salty taste, interpolated between levels 1 and 3 set in GOST 57164 (2016).

The assigned categories of the intensity of odor and taste in the prepared items were set according to the preparation procedure (ISO 17043 2010). The influence of any lack of homogeneity of the initial solutions on the assigned categories was negligible. The solutions of sodium hypochlorite and sodium sulfide were stable for 3 weeks when kept in tightly closed glassware between temperatures from 4 to 20 °C. The stability of the test items 1 and 2 was not relevant, as they were prepared immediately before use. The assigned category of the salty taste intensity was stable for 3 weeks when item 3 was kept in

tightly closed glassware between temperatures from 4 to 20 °C. Within-laboratory variability of 12 replicates studied at UNIIM did not exceed a deviation of one intensity level from the assigned category for chlorine or sulfurous odor at 20 and 60 °C, or salty taste.

The components of items 1 and 2, as well as item 3, were distributed to the 49 laboratories that participated in the comparison in random order. The laboratories received and examined the items within 5–10 days from the preparation of the solutions at UNIIM.

Methods of examination

Testers (technicians) having symptoms such as runny nose, allergic reactions, or headache were excluded from the test. The examination of the items was performed at a participating laboratory immediately after the preparation of the final solutions in the same conditions as for routine water samples. The methods of examination (GOST 57164 2016) are summarized below.

Examination of odor and its intensity at 20 and 60 °C

The temperature of a test item was measured and adjusted to 20 ± 2 °C by keeping it at room temperature in tightly closed glassware. About 100 cm^3 of the item was transferred into a glass-stoppered flask of 250–350 cm^3 and homogenized with rotating movements. Then, the flask was opened, and the odor and its intensity were examined.

To adjust a test item's temperature to 60 ± 5 °C, about 100 cm^3 of the item were transferred into a flask of 250–350 cm^3 closed by a watch glass. The flask was immersed in a water bath for heating. When the target temperature was achieved, the water was homogenized with rotating movements, the watch glass was removed, and the odor and its intensity were quickly examined.

Examination of taste and its intensity

About 30 cm^3 of the test item were taken into the oral cavity in small portions (about 15 cm^3), without swallowing, hold for 3–5 s and spat out. The time between the examination of two samples was not less than 30 s.

Examination responses

Each laboratory provided one result, i.e., one set of the expert examination responses presented in the Electronic Supplementary Material to this paper (RawData_Interlab_comp.pdf file).

Laboratory 3 did not report on the odor intensity at 60 °C; laboratory 18 did not report on the kind of odor; laboratories 38 and 39 did not report on the odor at all. Therefore, the responses of the remaining 45 laboratories were taken into account for analysis of the odor intensity.

Similar situations happened when examining the taste intensity. Laboratory 10 reported on the kind of taste mistakenly; laboratories 18, 19, and 37 did not report on the taste at all. Thus, the responses of the remaining 45 from 49 laboratories were taken into account for analysis of the taste intensity.

RESULTS AND DISCUSSION

Odor intensity

For the ORDANOVA model, there are: factor X1 – laboratory with $I = 45$ levels; factor X2 – temperature with $J = 2$ levels; $K = 6$ categories/levels of chlorine and sulfurous odor intensity; $n = 1$ – one examination response from each laboratory; $N = 90$ responses in total. The frequencies of the responses from the RawData_Interlab_comp.pdf file are shown in Table 1 by categories and temperatures.

Two-way ORDANOVA without replication

The vectors of statistical sample relative frequencies of the responses by categories in Table 1 for chlorine and sulfurous odor intensity at the two temperatures are $\hat{\boldsymbol{p}} = (4/90, 38/90, 29/90, 19/90, 0, 0)$ and $\hat{\boldsymbol{p}} = (0, 0, 0, 22/90, 28/90, 40/90)$, respectively. The sample cumulative relative frequency vectors for chlorine and sulfurous odor intensity are $\hat{\boldsymbol{F}} = (4/90, 42/90, 71/90, 1, 1, 1)$ and $\hat{\boldsymbol{F}} = (0, 0, 0, 22/90, 50/90, 1)$, respectively. The total sample variation of the responses for the intensity of chlorine odor is $\hat{V}_T = 0.366$, and for sulfurous odor it is $\hat{V}_T = 0.345$ with $df_T = 89$ by Equation (3). The between-laboratory variation for the intensity of chlorine odor is $\hat{C}_B = 0.256$, and for sulfurous odor it is $\hat{C}_B = 0.250$ with $df_B = 45$ by Equation (5). The residual variation for the intensity of chlorine odor is $\hat{V}_W = 0.110$, while for sulfurous odor it is $\hat{V}_W = 0.096$ with $df_W = 44$ by Equation (6). The fraction of the total variation reflecting the between-laboratory effect on the response for the intensity of chlorine odor is

Table 1 | Frequencies of the responses of chlorine and sulfurous odor intensity

	Frequency			
	Chlorine		Sulfurous	
Category	20 °C	60 °C	20 °C	60 °C
0	2	2	0	0
1	24	14	0	0
2	10	19	0	0
3	9	10	10	12
4	0	0	17	11
5	0	0	18	22

$\hat{R}^2_B = 0.700$, and for sulfurous odor it is $\hat{R}^2_B = 0.725$ by Equation (10). This indicates that there is a joint influence of a laboratory and temperature on the variability of chlorine and sulfurous odor intensity responses by categories. However, the fractions $\hat{R}^2_{X1} = 0.670$ and $\hat{R}^2_{X2} = 0.030$ for chlorine intensity, and similarly, $\hat{R}^2_{X1} = 0.717$ and $\hat{R}^2_{X2} = 0.006$ for sulfurous odor intensity by Equation (11) show that a laboratory is a good predictor of odor intensity, whereas temperature impacts the responses much less.

Table 2 details the decomposition of the total sample variation by laboratory (factor $X1$) and temperature (factor $X2$), the significance indices $\widehat{SI}$ by Equation (13), and the critical values SI_{crit}, evaluated using the calculator tool at 95 % level of confidence and 10,000 Monte Carlo trials.

The significance index of the laboratory factor $\widehat{SI}_{X1} = 1.360$ for chlorine odor intensity exceeds its critical value of 1.185 at 95% level of confidence; similarly for sulfurous odor intensity $\widehat{SI}_{X1} = 1.454$ exceeds its critical value of 1.202. At the same time, the significance index of the temperature factor does not exceed its critical value at 95% level of confidence for both chlorine odor intensity ($\widehat{SI}_{X2} = 2.423 < 3.010$) and sulfurous odor intensity ($\widehat{SI}_{X2} = 0.511 < 3.248$). That means rejecting the null hypothesis concerning the (zero) difference between laboratories in classifying chlorine or sulfurous odor intensity by categories/levels: this difference is statistically significant. The effect of temperature in classifying chlorine or sulfurous odor intensity by categories is not significant as the null hypothesis is not rejected. Similar insignificance of temperature was reported in the thesis of Whelton (2001) for isobutanal in drinking water, while the perception of some other odorants was affected by temperature changes from 25 to 45 °C. Note that this effect might depend on the odorant concentration in water.

The simulated distributions of the two significance indices are presented in Figure 1. The critical values SI_{crit} for 95 % level of confidence in Table 2 correspond to $\widehat{SI}$ values in Figure 1 when 95 % of the area under the curve of relative frequency is achieved.

Decomposition of the between-laboratory variation component by Equation (9) according to the categories of the obtained chlorine odor intensity responses $k = 0$, 1, 2, 3, 4, and 5 leads to $\hat{C}_B(0) = 0.020$, $\hat{C}_B(1) = 0.172$, $\hat{C}_B(2) = 0.128$, $\hat{C}_B(3) = 0$, $\hat{C}_B(4) = 0$, and $\hat{C}_B(5) = 0$. This means that capabilities of the laboratories to identify chlorine odor intensity are better (dispersions of the responses are smaller) for categories $k = 0$, 3, 4, and 5 than for categories $k =$ 1 and 2. It seems strange that the capabilities to identify the odor intensity of categories $k = 3$, 4, and 5 are assessed as perfect, while no response fell into these categories. However, this is due to the fact that the testers of all the laboratories found correctly that the item odor intensity does not belonging to those categories. Therefore, also the cumulative frequency achieved 1 at $k = 3$.

Table 2 | Results of the two-way ORDANOVA without replicates for the chlorine and sulfurous odor intensity responses

Odor	Factor	Variation component	$\hat{R}^2$	$\widehat{SI}$	df	SI_{crit}
Chlorine	Laboratory ($X1$)	0.246	0.670	1.360	44	1.185
	Temperature ($X2$)	0.010	0.030	2.423	1	3.010
Sulfurous	Laboratory ($X1$)	0.248	0.717	1.454	44	1.202
	Temperature ($X2$)	0.002	0.006	0.511	1	3.248

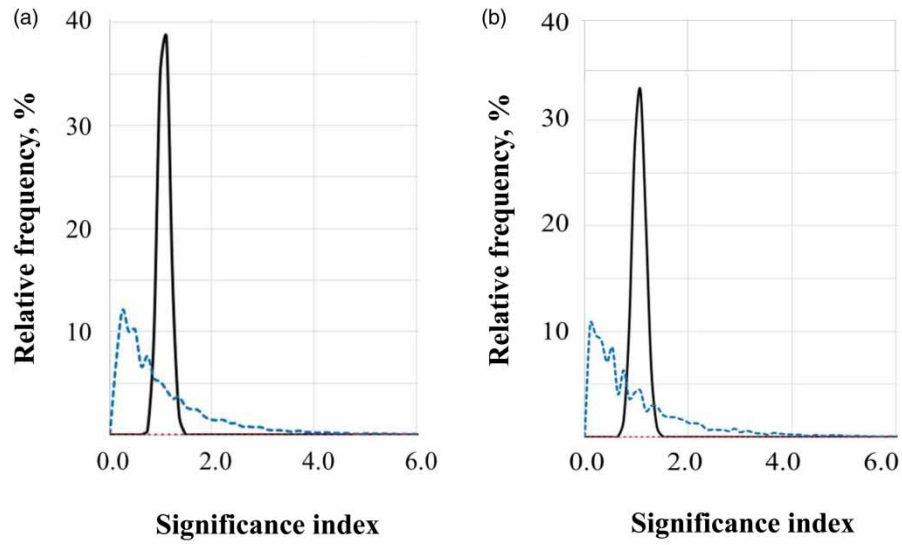


Figure 1 | Empirical distribution functions of $\widehat{SI}_{X1}$ (solid black line) and $\widehat{SI}_{X2}$ (dashed blue line) for chlorine (a) and sulfurous (b) odor intensity. The significance index of the factor interaction $\widehat{SI}_{X1*X2}$ (dashed red line) is not applicable and hence equal to zero in the plot. Please refer to the online version of this paper to see this figure in colour: http://dx.doi.org/10.2166/wh.2022.060.

Similarly, for sulfurous odor intensity, $\hat{C}_B(0) = 0$, $\hat{C}_B(1) = 0$, $\hat{C}_B(2) = 0$ and $\hat{C}_B(3) = 0.141$, $\hat{C}_B(4) = 0.171$, $\hat{C}_B(5) = 0$. Thus, capabilities of the laboratories to identify sulfurous odor intensity are better for categories $k = 0, 1, 2$ than for categories $k = 3$ and 4 as no response fell into categories $k = 0, 1, 2$. The discussed variation for category $k = 5$ equals zero by definition of Equation (9).

Note that the consensus (ISO 17043 2010) of the laboratories for the chlorine odor intensity, being defined here as the most frequent response level/category at both 20 and 60 °C, is $k = 1$ ($\hat{p}_{..1} = 38/90$), whereas the assigned level was $k = 2$ ($\hat{p}_{..2} = 29/90$). The consensus of the laboratories for the sulfurous odor intensity is $k = 5$ ($\hat{p}_{..5} = 40/90$), whereas the assigned level was $k = 4$ ($\hat{p}_{..4} = 28/90$). Detecting the odor intensity levels for those categories that are far from the assigned category is simple, e.g., for chlorine intensity there was no response related to categories $k = 4$ and 5 as the assigned category was 2. For sulfurous intensity, there was no response in categories $k = 0$, 1, and 2 as the assigned category was 4. In both cases, the responses were mostly around ± 1 of the assigned category.

The UNIIM decision was to accept a deviation of a laboratory response from the assigned category for one level as satisfactory. However, in general, deviation of the consensus of the 45 laboratories from the assigned category requires an additional analysis of the inconsistency, training of the technicians (testers), and a repetition of the interlaboratory comparison.

Comparison with the two-way ANOVA without replication

The data in the RawData_Interlab_comp.pdf file for chlorine and sulfurous odor intensity were analyzed with Excel using the two-way ANOVA, treating the intensity response as a continuous variable. The results of the analysis are given in Table 3,

Table 3 | Results of two-way ANOVA without replicates for the chlorine and sulfurous odor intensity responses

Odor	Source of variation	SS	df	MS	F	P-value	F_{crit}
Chlorine	Laboratories	48.400	44	1.100	3.194	0.000	1.651
	Temperatures	1.344	1	1.344	3.903	0.054	4.062
	Error	15.156	44	0.344			
	Total	64.900	89				
Sulfurous	Laboratories	42.289	44	0.961	2.730	0.001	1.651
	Temperatures	0.011	1	0.011	0.032	0.860	4.062
	Error	15.489	44	0.352			
	Total	57.789	89				

where SS is the sum of squares of deviations of responses from the average value; df is the number of degrees of freedom; $MS = SS/df$ is the mean deviation square; F is the empirical value of the Fisher criterion; P-value is the minimal probability of rejecting the null hypothesis on homogeneity when it is correct; and F_{crit} is the critical value of F at 95 % confidence level.

From Table 3, it follows that rejecting the null hypothesis about the homogeneity of the laboratory responses at 95% level of confidence, i.e., the differences between the responses of the laboratories are significant. The null hypothesis about a (zero) difference between responses obtained at 20 and 60 °C is not rejected at 95% level of confidence – there is not a significant difference between the two levels of temperature. Thus, the results of the testing significance of the effects based on the ORDANOVA and ANOVA models in this experiment are in agreement.

Intensity of salty taste

There is one factor – laboratory with $I = 45$ levels; $K = 6$ categories/levels of taste intensity; $n = 1$ – one examination response from each laboratory; and $N = 45$ responses in total. Frequencies of the responses from the RawData_Interlab_comp.pdf file by the categories discussed below.

One-way ORDANOVA

The vector of sample relative frequencies is $\hat{\boldsymbol{p}} = (0, 14/45, 16/45, 15/45, 0, 0)$, and the vector of sample cumulative relative frequencies is $\hat{\boldsymbol{F}} = (0, 14/45, 30/45, 1, 1, 1)$. The total sample variation of the responses is $\hat{V}_T = 0.349$ with $df_T = 44$ by the formula derived from Equation (3) for one response from each laboratory, whereas the within-laboratory component is $\hat{V}_W = 0$. The between-laboratory variation is $\hat{C}_B = 0.349$ with $df_B = 44$ by the decomposition theorem. As between-laboratory variation and the total variation coincide in this case, $\hat{R}_B^2 = 1$ by Equation (10) indicates a perfect predictability of a laboratory response on taste intensity by categories. Moreover, the significance index is $\widehat{SI} = 1$ by Equation (13); hence, the null hypothesis of homogeneity between laboratories is not rejected, i.e., the responses of the laboratories on the salty taste intensity are not statistically different.

Decomposition of the between-variation component by categories by Equation (9) leads to the following: $\hat{C}_B(0) = 0$, $\hat{C}_B(1) = 0.214$, $\hat{C}_B(2) = 0.222$, $\hat{C}_B(3) = 0$, $\hat{C}_B(4) = 0$, and $\hat{C}_B(5) = 0$. This means that capabilities of the laboratories to identify salty taste intensity are better for categories $k = 0$, 3, 4, and 5 than for categories $k = 1$ and 2 (no response fell into categories $k = 0$, 4 and 5, and the cumulative frequencies achieved 1 at $k = 3$).

The consensus of the laboratories is the salty taste intensity category $k = 2$ ($\hat{p}_{.2} = 16/45$), coinciding with the assigned level.

Comparison with the one-way ANOVA

The data in the RawData_Interlab_comp.pdf file for salty taste intensity were analyzed with Excel using the one-way ANOVA, treating the salty taste intensity as a continuous variable. The total sum of squares $SS_T = 28.978$ with $df_T = 44$ (for one response from each laboratory) is equal to the between-laboratory sum $SS_B = 28.978$ with $df_B = 44$, and the mean square is $MS_B = 0.659$. As the within-laboratory variation is not evaluated in the absence of replicates, the Fisher criterion is not formally applicable here.

However, using the UNIIM maximum within-laboratory deviation of 12 replicate responses from the assigned category for 1 level/category as an approximation, the maximum within-laboratory sum of squares can be assumed equal to $SS_W = 1$ with $df_W = 11$. Hence, the mean square $MS_W = 0.091$, and the minimum empirical value of the Fisher criterion can be simulated as $F = MS_B/MS_W = 7.242$. As the critical value for $df_B = 44$ and $df_W = 11$ is $F_{crit} = 2.520$ at 95% level of confidence, and the P-value is 0.001, the null hypothesis of homogeneity of the responses is rejected. Thus, the responses of the laboratories on the salty taste intensity differed statistically, even assuming the extremely large approximation of the variation MS_W.

Note that, in contrast to the case of examination of the odor intensity, the results of analysis of the taste intensity responses with the two methods, ANOVA and ORDANOVA, are in contradiction. In such a case, it is clear that ORDANOVA results are reliable, while ANOVA, performed with the violation of its basic assumptions, may lead to mistaken results.

CONCLUSIONS

The two-way ORDANOVA without replication was applied for the first time to an interlaboratory comparison of ordinal data from a human organoleptic examination of the intensity of odor of drinking water, which is performed at 49 ecological laboratories. Using a decomposition of the total variation of the ordinal data and simulation of the multinomial distribution of the relative frequencies of the data in different categories, the statistical significance of the interlaboratory variation of the

laboratories' responses for both chlorine and sulfurous odor intensity was shown. No influence of the temperature (20 and 60 °C) of test items on the responses was detected. This effect may depend on the chemical properties of the odorants and their concentrations in water. The statistical decomposition also allowed evaluation of the capability of the laboratories to identify different categories of odor intensity. It is noted that the consensus (the most frequent) response of the laboratories differed from the assigned category by only one level on the ordinal scale.

The one-way ORDANOVA was used for the analysis of salty taste intensity, where the interlaboratory variability was found to be statistically insignificant. The capability of the laboratories to identify different categories was also evaluated. The consensus response of the laboratories for the taste intensity coincided with the assigned category.

A comparison of ORDANOVA and ANOVA results showed that ORDANOVA provides a more useful tool for ordinal data. Concerning the statistical significance of the effects, the results of both the methods may, in general, be the same or different. However, when ANOVA is applied for categorical data, its basic assumption of additivity of variables is violated, and so the results obtained cannot be trusted.

ACKNOWLEDGEMENTS

The authors would like to thank O.N. Kremleva and E.V. Rudnitskaya, UNIIM, Russia, for participating in organization of the interlaboratory comparison.

AUTHOR CONTRIBUTION

T.G. developed the methodology, conducted a formal analysis, did software analysis, visualized the article, and wrote the review and edited the article. I.K. conceptualized the whole article, administered the project, visualized the article, and wrote the review and edited the article. F.P. and D.B.H. wrote the review and edited the article. A.A.S. and P.S.C. validated the article. V.N.N found the resources.

FUNDING

This research was supported, in part, by the International Union of Pure and Applied Chemistry (Project 2021-017-2-500).

CONFLICT OF INTEREST

The authors declare no competing interests.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Agresti, A. 2012 *Categorical Data Analysis*, 3rd edn. Wiley, New Jersey.

Baird, R. B., Eaton, A. D. & Rice, E. W. 2018 *Standard Methods for the Examination of Water and Wastewater*, 23rd edn. Parts 2150 Odor, 2160 Taste, 2170 Flavor Profile Analysis. APHA, AWWA, WEF, Washington.

Bashkansky, E., Gadrich, T. & Kuselman, I. 2012 Interlaboratory comparison of test results of an ordinal or nominal binary property: analysis of variation. *Accredit. Qual. Assur.* **17**, 239–243. https://doi.org/10.1007/s00769-011-0856-0.

Burlingame, G. A., Doty, R. L. & Dietrich, A. M. 2017 Humans as sensors to evaluate drinking water taste and odor: a review. *J. Am. Water Works Assoc.* **109**, 13–24. https://doi.org/10.5942/jawwa.2017.109.0118.

da Silva, R. B. & Ellison, S. L. R. 2021 *Eurachem/CITAC Guide: Assessment of Performance and Uncertainty in Qualitative Chemical Analysis*. Available from: https://www.eurachem.org (accessed 11 January 2022).

Gadrich, T. & Bashkansky, E. 2012 ORDANOVA: analysis of ordinal variation. *J. Stat. Plann. Inference* **142**, 3174–3188. https://doi.org/10.1016/j.jspi.2012.06.004.

Gadrich, T. & Marmor, Y. N. 2021 Two-way ORDANOVA: analyzing ordinal variation in a cross-balanced design. *J. Stat. Plann. Inference* **215**, 330–343. https://doi.org/10.1016/j.jspi.2021.04.005.

Gadrich, T., Bashkansky, E. & Kuselman, I. 2013 Comparison of biased and unbiased estimators of variances of qualitative and semi-quantitative results of testing. *Accredit. Qual. Assur.* **18**, 85–90. https://doi.org/10.1007/s00769-012-0939-6.

Gadrich, T., Kuselman, I. & Andrić, I. 2020 Macroscopic examination of welds: interlaboratory comparison of nominal data. *SN Appl. Sci.* **2**, 2168. https://doi.org/10.1007/s42452-020-03907-4.

GOST R 57164 2016 *Drinking Water. Methods for Determination of Odor, Taste and Turbidity*. Available from: https://runorm.com/catalog/1004/876961/ (accessed 11 January 2022).

Hibbert, D. B. 2020 Chemometric analysis of sensory data. In: *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis* (Brown, S., Tauler, R. & Walczak, B., eds.). Elsevier, pp. 149–192. https://doi.org/10.1016/B978-044452701-1.00010-7.

Hibbert, D. B., Korte, E.-H. & Örnemark, U. 2021 Fundamental and metrological concepts in analytical chemistry (IUPAC Recommendations 2021). *Pure Appl. Chem.* **93**, 997–1048. https://doi.org/10.1515/pac-2019-0819.

ISO 8586 2014 *Sensory Analysis. General Guidelines for the Selection, Training and Monitoring of the Selected Assessors and Expert Sensory Assessors*. Available from: https://www.iso.org/standard/45352.html (accessed 11 January 2022).

ISO/IEC 17043 2010 *Conformity Assessment – General Requirements for Proficiency Testing*. Available from: https://www.iso.org/standard/29366.html (accessed 11 January 2022).

ISO/IEC 17025 2017 *General Requirements for the Competence of Testing and Calibration Laboratories*. Available from: https://www.iso.org/standard/66912.html (accessed 11 January 2022).

ISO/TS 20612 2007 *Water Quality. Interlaboratory Comparisons for Proficiency Testing of Analytical Chemistry Laboratories*. Available from: https://www.iso.org/standard/46269.html (accessed 11 January 2022).

Kaloudis, T., Dietrich, A., Zamyadi, A., Lin, T.-F. & Lado, R. 2021 Water taste and odour (T&O): challenges, gaps and solutions. CEJ Adv. Available from: https://www.sciencedirect.com/journal/chemical-engineering-journal-advances/special-issue/10D4KDNCQFQ (accessed 4 April 2022).

Light, R. J. & Margolin, B. H. 1971 An analysis of variance for categorical data. *J. Am. Stat. Assoc.* **66**, 534–544. https://doi.org/10.1080/01621459.1971.10482297.

Lin, T. F., Watson, S. & (Mel) Suffet, I. H. 2019 *Taste and Odour in Source and Drinking Water: Causes, Controls, and Consequences*. IWA Publishing, London.

Magnusson, B. & Ornemark, U. 2014 *Eurachem Guide: The Fitness for Purpose of Analytical Methods – A Laboratory Guide to Method Validation and Related Topics*. Available from: https://www.eurachem.org (accessed 15 February 2022).

Marmor, Y. N. & Gadrich, T. 2019 *Two-Way ORDANOVA Tool. V1*. Available from: https://docs.google.com/spreadsheets/d/1fSF8ZpZf0pNRRm9l_dh4fN14nyLDROms/edit?usp=sharing&ouid=101198859611382942102&rtpof=true&sd=true (accessed 21 November 2021).

NIST/SEMATECH 2021 *e-Handbook of Statistical Methods*. Available from: https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/multpdf.htm (accessed 1 December 2021).

Scheffé, H. 1999 *Analysis of Variance*. Willey Classics Library, New York.

Tiikkainen, U., Ciaralli, L., Laurent, C., Obkircher, M., Patriarca, M., Robouch, P. & Sarkany, E. 2022 Is harmonization of performance assessment in non-quantitative proficiency testing possible/necessary? *Accredit. Qual. Assur.* **27**, 1–8. https://doi.org/10.1007/s00769-021-01492-6.

Whelton, A. J. 2001 *Temperature Effects on Drinking Water Odor Perception. Thesis of Master of Science in Environmental Engineering*, Virginia Polytechnic Institute and State University. Available from: https://vtechworks.lib.vt.edu/handle/10919/36221 (accessed 6 April 2022).

First received 18 February 2022; accepted in revised form 14 May 2022. Available online 26 May 2022