Accreditation and Quality Assurance (2022) 27:55–83
https://doi.org/10.1007/s00769-021-01487-3

GENERAL PAPER

Trends in the development of proficiency testing for chemical analysis: focus on food and environmental matrices

Igor Renato Bertoni Olivares[1] · Gilberto Batista de Souza[2] · Ana Rita de Araujo Nogueira[2] · Vitor Hugo Polisél Pacces[1] · Pamela Aparecida Grizotto[1] · Paula Souza da Silva Gomes Lima[1] · Rhaissa Mecca Bontempi[1]

Received: 2 February 2021 / Accepted: 15 November 2021 / Published online: 29 January 2022

Abstract
The importance of quality in analytical chemistry stimulates the development of different tools to assure the reliability of analytical results. Among different tools, proficiency testing (PT) stands out because it can be used to evaluate bias, check uncertainty, train analysts, or certify if a laboratory can execute a method adequately and provide correct results. There is a growing demand for traceable and reliable results in analytical chemistry, which can be illustrated with the growth of ISO/IEC 17025 accreditation and the importance of PT in this context. This has led to an increase in developments and publications about PT programs. This paper reports a detailed review considering the best practices to develop PT for chemical analysis, focusing on food and environmental matrices. An evaluation of the trends and the statistical strategies in its development in the last two years was performed to guide new developments of this tool that is increasingly necessary for laboratories.

Keywords Proficiency testing · Quality assurance (QA) · Uncertainty · Stability · Homogeneity · Value assignment

Introduction

The reliability of analytical results has always been a concern of laboratories, a fact highlighted more than 20 years ago by Valcarcel [1]. This concern has led to the development of new concepts and tools [2, 3], also stimulating the development and revision of many standards applicable to quality management systems for laboratories, such as ISO/IEC 17025 [4], GLP [5], and ISO 15189 [6].

Due to the increasing need to provide reliable results, different regulatory agencies require implementing quality management systems by laboratories, promoting growth in the number of accreditations in the use of different standards, such as ISO/IEC 17025 (Fig. 1).

As shown in Fig. 1, the application of quality management in laboratories is a global trend, increasing the need to develop new knowledge in areas like validation, uncertainty estimation, control charts, production of certified reference materials (CRM), development of proficiency testing (PT), application of statistics, and laboratory management software, leading to the updating and development of technical standards and legislation supporting these concepts [8].

The analytical quality assurance cycle (AQAC) is a proposal for a specific quality tool for laboratories, involving three concepts considered essential: method validation, uncertainty estimation, and quality control (QC).

The AQAC [2] is a conceptual tool that correlates the three most important requirements, generally applied in an ISO/IEC 17025 accredited laboratory, that are essential to provide reliable results. Considering that these three requirements are supported by statistical concepts, it is possible to note that there is an important correlation between them. The AQAC starts with validation to demonstrate that the method is fit for the purpose. The second step is the measurement of uncertainty to evaluate the dispersion of the quantity values being attributed to a measurand (generally, the principal uncertainties sources are from validation, like precision or linearity). The third step is the quality control (QC) that provides ongoing validation of the method. Figure 2 shows the AQAC with the insertion of CRM and PT

✉ Igor Renato Bertoni Olivares
igorolivares@hotmail.com

[1] Chemistry Institute of São Carlos (IQSC-USP), University of São Paulo, Av. Trabalhador São-carlense, 400 CP 780, São Carlos, SP CEP 13560-970, Brazil

[2] Embrapa Pecuária Sudeste, Rodovia Washington Luiz, km 234, CP 339, São Carlos, SP CEP 13560-970, Brazil

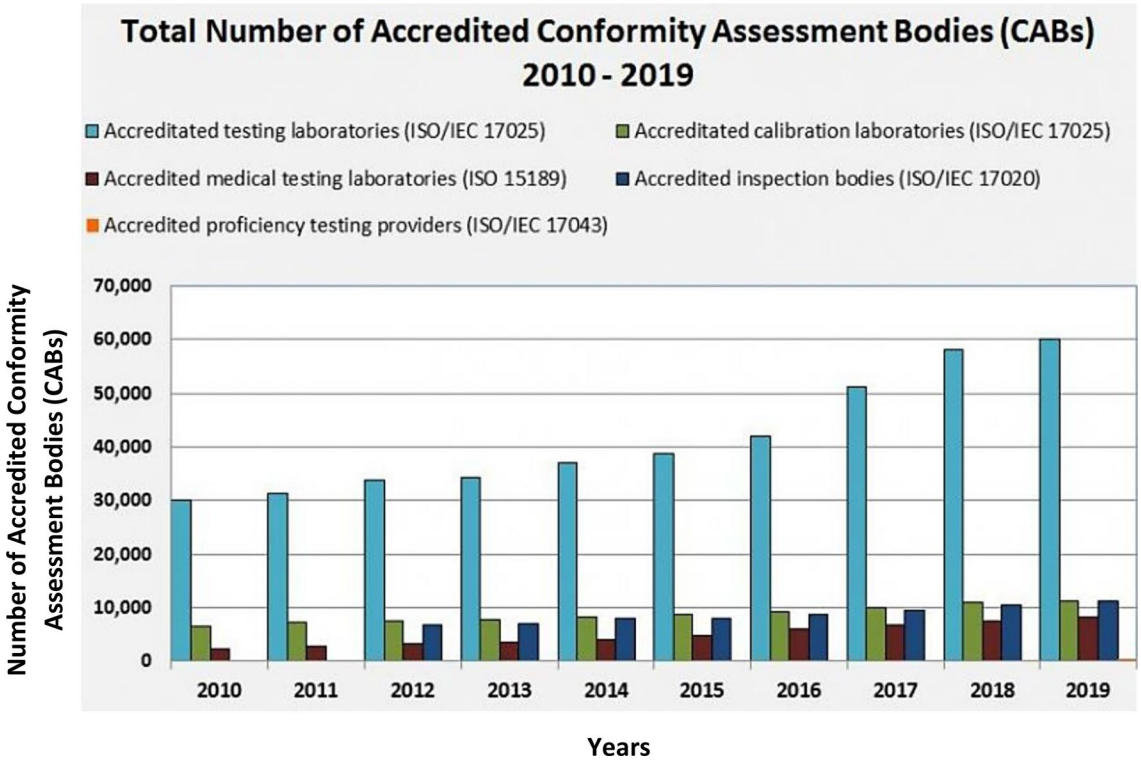

Fig. 1 Accredited laboratories considering the accreditation bodies from approximately 114 countries that are ILAC members and associates. Data obtained from ILAC (ILAC is the international authority on laboratory and inspection body accreditation, with a membership consisting of accreditation bodies and stakeholder organizations worldwide) [7]

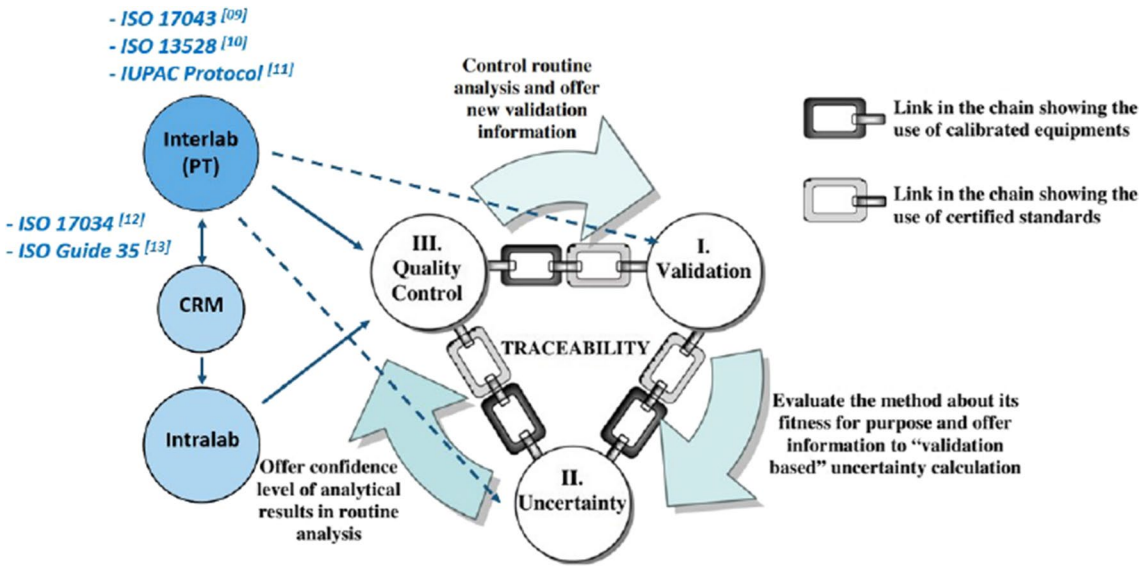

Fig. 2 Analytical quality assurance cycle (AQAC) adapted with the CRM and PT concepts

concepts, indicating that they are closely linked to this tool and are necessary for this quality cycle.

Figure 2 highlights that PT is a tool correlated with all the concepts of this cycle. This way, the results of PT can be used:

- to evaluate the performance of laboratories for specific tests or measurements;
- to monitor laboratories' continuing performance; and
- to evaluate if the estimated measurement uncertainty of the result was underestimated or overestimated by comparing the result obtained by the laboratory and its uncertainty with the assigned value of the PT sample and its acceptance range.

The importance of participation of PT is also highlighted in the quality management systems for laboratories. This periodic participation is mandatory for laboratories that want accreditation (formal recognition) to ISO/IEC 17025 by an accreditation body [9], which increases demand and development of new PT, generating an increase in the volume of publications on this subject.

The development and application of PT involve a series of steps, which despite being harmonized by standards such as ISO 13528 [10] and ISO/IEC 17043 [11], include different approaches, which must be chosen depending on the matrix and analytes that will be evaluated by PT, such as assigning values obtained by consensus or a reference value (which can be calculated by different strategies); different ways to performance evaluation (z-score, Zeta-score, etc.); use of different graphical methods (Youden plot, histograms, etc.); and different strategies for stability and homogeneity evaluation of PT items, among others.

Therefore, considering the increasing number of accredited laboratories and corresponding rising demand for PT, combined with the different strategies regarding its development and application, this review article presents a bibliographic review of published articles, focusing on food and environmental matrices. In this respect, we emphasize the trends in PT developments for chemical analysis, seeking to assist those who intend to develop new PT protocols or want to know more about the theme and the main approaches currently applied.

Discussion

Search method

Three different searches were carried out to obtain an overview of PT publications, as illustrated in Fig. 3. The first search for papers used the expression "proficiency testing"

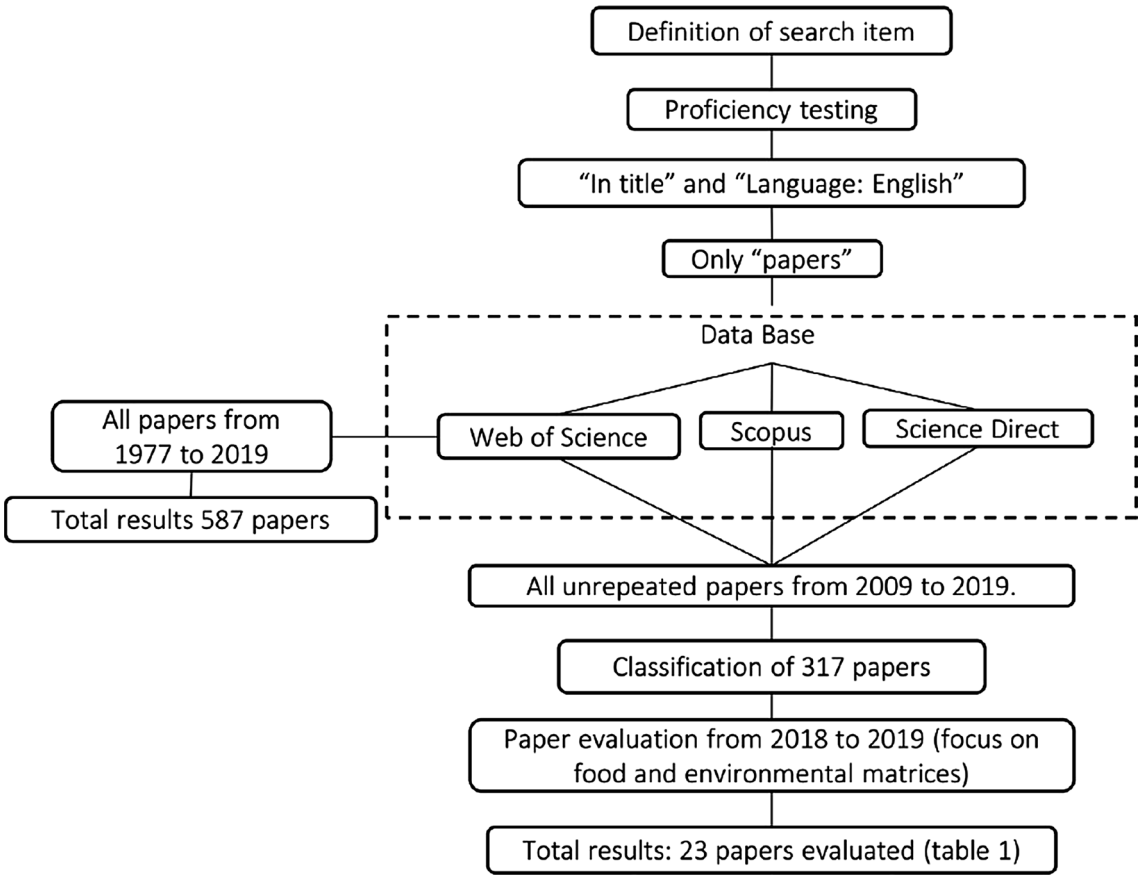

Fig. 3 Method used for the search and exclusion of papers

in the title, considered only published papers not repeated in the last 10 years (2009–2019), obtaining a total of 317 papers. Three databases were used, Web of Science, Scopus, and Science Direct.

Since Web of Science generated the most results, we performed a new search only in this database, considering all years with the expression "proficiency testing" contained in the title. This search resulted in 587 non-repeated papers published between 1977 and 2019.

Considering the necessity to select a sample to do an evaluation of the trends and the statistical strategies in PT development for chemical analysis, a third search was carried out in the years 2018–2019, including all evaluated databases focused on food and environmental matrices, considering that during 10 years (2010–2019) 75% of the PT was focused on these two areas of study. The number of non-repeated papers found was 23 quantitative PTs. Each paper was evaluated, considering the main steps of PT. The results are shown in Table 1.

Arguments concerning these steps are detailed in this paper.

In general, the papers were "theoretical papers" or "PT applications." Papers classified as "theoretical" focus on issues such as the best way to calculate and interpret results and compare techniques. In most circumstances, theoretical PT papers promote discussion using data from PT performed during different time intervals with different matrices. Papers classified as "application" generally describe PT planning and organization and discuss participating laboratories' performance.

Observed trends

The analysis of the data obtained by the flowchart in Fig. 3 revealed several trends. Since the database that provided the most results was Web of Science, we decided to check the frequency of publications over the years in this database (Fig. 4). A significant increase in publications was noted since the 2000s.

According to the study by Olivares et al. [35], it is possible to observe a growing interest in CRM research and development from 1976 to 2016, intensifying significantly since 1996, mainly after 2006. The volume of PT studies followed the same behavior, as can be seen in Fig. 4. This correlation is explained by the requirement for participation in PT of laboratories with accreditation according to ISO/IEC 17025, an international standard implemented in 2000. Another fact that could explain these numbers shown in Fig. 4 is that the accreditation of PT providers also predominantly started in the year 2000, with the publication of some ILAC documents, and with the publication of ISO/IEC 17043 [11] in 2010, which accelerated the accreditation of PT providers even further.

The 317 articles found between 2009 and 2019 were mostly about quantitative PTs, with really few papers about qualitative or interpretative results. These papers were focused mainly on three research areas: physics, chemistry, and biology, and thus could be separated according to the nature of the PT (execution or theoretical) reported (Fig. 5). In these research areas, the matrices submitted to PT were contained in five main categories: (1) food (meat, fruits, animal feed, beverages, cereals, dairy products, oils); (2) environmental (soil, waste, air, water); (3) biological material (blood, urine, serum, muscle tissue, hair, organs, microorganisms, saliva); (4) inorganic (metal alloys, polymers, radioactive materials, textiles, ceramics, oil); and (5) instrumentation associated with clinical trials (thermometers, manometers). Figure 6 shows a new division considering only the "execution" and the matrices evaluated. It shows that 75% of the PT was focused on two main areas of study: food and environmental. This trend is intuitive because these are precisely the areas considered strategic by most countries. For this reason, testing in these areas is under more significant pressure due to government inspections and accreditation requirements, mainly regarding ISO/IEC 17025. To ensure the validity of results, the standard requires the laboratory to monitor its performance by participating in PT or any other type of interlaboratory comparison [4]. Because of this trend, PT papers involving food and environmental matrices are the focus of this study.

There are several types of PT providers: government agencies, universities, and private companies. Regardless of the origin, the discussions focus on the providers' evaluation parameters, such as development of the PT items, stability, homogeneity, attribution of assigned values, performance evaluation, and several graphical methods for interpretation of results. From the data in Fig. 5, it can be noted that the chemical area had the most published papers, indicating the interest in developing PT in this area. The global impact can be explained by the increasing adherence of chemical testing laboratories to ISO/IEC 17025 accreditation (Fig. 1).

The European PT Information System (EPTIS) database gathers records of PT schemes regularly applied. It started in 2000 and currently there are 4,445 PT schemes registered in EPTIS by 52 collaborating countries [36]. It is possible to highlight some areas in this database, such as "Food and Drink" with 12.1 % of the total, and "Environmental (water, soil, and sludge)" with 10.6 %.

The concern with analytical results' quality is growing, but perhaps still not enough. Monya Baker [37] pointed out a worrying trend: in a general context, the scientific community recognizes that currently there is a crisis of reproducibility of published experiments, particularly in the areas of chemistry, biology, physics, and engineering. Several factors were pointed out by the interviewed researchers as causes of this crisis, but we draw attention to these:

Table 1 Summarized results from papers on Proficiency Testing, published in the last two years (2018–2019), focusing on food and environmental matrices

References	Article title	Number of participants	Matrix	Analytes	Is this PT registered with EPTIS? (www.eptis.org)	Is there a PT in EPTIS for these analytes involving this matrix? If so, which one
[12]	31P NMR Method for Phospholipid Analysis in Krill Oil: Proficiency Testing—A Step toward Becoming an Official Method	17	Krill Oil	5 phospholipid (PL) species + total P	No	No
[13]	A review of the TAEA proficiency test on natural and anthropogenic radionuclides activities in black tea	20	Black tea powder mixed with contaminated tea that was withdrawn from the market after the Chernobyl accident	^{137}Cs, ^{40}K and ^{90}Sr massic activities	No	Yes https://www.eptis.bam.de/eptis/WebSearch/view/557637—water https://www.eptis.bam.de/eptis/WebSearch/view/450013—soybean and grain
[14]	Comparison of Assigned Values from Participants' Results, Spiked Concentrations of Test Samples, and Isotope Dilution Mass Spectrometric Results in Proficiency Testing for Pesticide Residue Analysis	113–119	Vegetable-paste of corn, pumpkin and spinach	organophosphorus pesticides	No	Yes https://www.eptis.bam.de/eptis/WebSearch/view/557179—pasta de tomate https://www.eptis.bam.de/eptis/WebSearch/view/438732—organofosforados em alimentos originarios de plantas
[15]	Comparison of relative INAA and k0-INAA using proficiency test materials at ITU TRIGA Mark II research reactor	Not stated	PLANTS, like Leek, lucerne, oil palm leaves, grass, Tobacco, rice and others & SOILS, like Sandy soil, calcareous brown soil, clay, river clay	As, Ba, Br, Ce, Co, Cr, Cs, Fe, K, La, Mn, Na, Na, Rb, Sb, Sc, Th, U, Zn	No	Yes Soil https://www.eptis.bam.de/pts13 3380 Plants https://www.eptis.bam.de/pts13 3667
[16]	Production and characterization of a traceable NORM material and its use in proficiency testing of gamma-ray spectrometry laboratories	9	Backflush of a drinking water treatment facility	Naturally occurring radioactive materials (NORM)—^{40}K, ^{226}Ra and ^{228}Ra	No	Yes https://www.eptis.bam.de/eptis/WebSearch/view/557637
[17]	Proficiency test exercises for particulate systems at CTBT radionuclide laboratories	15 to 16	Air PT (air Filter)	Radionuclide	No	No
[18]	Scents in the stack: olfactometric proficiency testing with an emission simulation apparatus	38	Emission simulation apparatus a replica of an industrial chimney with 23 m height	Amyl acetate ("AAC"); a mixture of organic, aromatic solvents, *n*-butanol, pig odor mixture, (R)-(+)-limonene, tetrahydrothiophene	No	Yes https://www.eptis.bam.de/eptis/WebSearch/view/205073

Table 1 (continued)

References	Article title	Number of participants	Matrix	Analytes	Is this PT registered with EPTIS? (www.eptis.org)	Is there a PT in EPTIS for these analytes involving this matrix? If so, which one
[19]	Can official control laboratories quantify reliably fipronil in eggs? Evidence from a proficiency testing round	Eighty-six National Reference Laboratories (NRLs), and EU Official Control Laboratories (OCLs) from 22 EU Member States, Norway, Serbia and Albania participated in a total of 108 laboratories	Eggs	Sum of fipronil plus its metabolite fipronil sulfone, expressed as fipronil	No	Yes CHINA: China NIL Research Center for Proficiency Testing (ID: 410429): Determination of Fipronil residue in egg products. Holland: RIKILT Wageningen University & Research (ID: 547011): Fipronil in egg and chicken fat—Tested property (Fipronil and fipronil sulfone) CHINA: Analysis Capability Assessment System of Chinese Academy of Inspection and Quarantine (ID: 663818): Determination of fipronil and its metabolites residue in egg proficiency testing) Tested property—(fipronil sulfone) Argentina: ILT—Interlaboratory Test S.A. (ID: 521649) Pesticide determination in sesame seeds. Tested property—(Fipronil) Italy: Test Veritas S.r.l. (ID: 136333). Eggs. Progetto Trieste. Tested property (Fipronil) France: BIPEA Proficiency Testing (ID: 156465). Multi-residue screening of pesticides [EIL Pesticides: Multirésidus pesticides]. Tested property—(Fipronil)

Table 1 (continued)

References	Article title	Number of participants	Matrix	Analytes	Is this PT registered with EPTIS? (www.eptis.org)	Is there a PT in EPTIS for these analytes involving this matrix? If so, which one
[20]	Comparisons between reproducibility standard deviations (SDR) derived from proficiency tests and from collaborative trials: mycotoxins in food	Proficiency test results were taken from the FAPAS scheme [4] from the years 2009–2015. FAPAS is a mature scheme and accredited to ISO/IEC 17043. It is the world's largest and most inclusive scheme for foodstuffs, covering roughly 1400 analyte/test material combinations and with participants from over 140 countries	The results of food samples from the FAPAS proficiency tests between the years 2009 to 2015	Analytes examined here were: aflatoxins B1, B2, G1, and G2; ochratoxin A; zearalenone; and deoxynivalenol	Yes	Yes United Kingdom (ID: 140970): Fera Science Ltd. FAPAS Food Chemistry. Tested property Mycotoxin Contamination) (Mycotoxin Contamination) (Multi-mycotoxins) United States: (ID: 156706): AAFCO Proficiency Testing Program (Formerly the AAFCO Collaborative Check Sample Program)—Mycotoxin Contaminant Scheme [Mycotoxin Contaminant Scheme]. (ID: 293216). France: BIPEA Proficiency Testing. Mycotoxins [EIL Contaminants alimentaires—Mycotoxines. (PTS Food contaminants: 31—Mycotoxins) South Africa (ID:629405): National Metrology Institute of South Africa (NMISA). Mycotoxins in cassava/animal feed/maize. Czech Republic (ID: 168632): Central Institute for Supervising and Testing in Agriculture (UKZUZ), Department of Proficiency Testing Programmes. Mycotoxins in Feedstuff and Food [MPZ ÚKZÚZ Mykotoxiny] Germany (ID: 640468): DLA—Proficiency Tests GmbH. Mycotoxin-Screening: Aflatoxins, Ochratoxin A, Deoxynivalenol, Zearalenon and Fumonisins in Breakfast Cereals (Muesli with corn and/or other cereal products, almonds and dried fruits)

Table 1 (continued)

References	Article title	Number of participants	Matrix	Analytes	Is this PT registered with EPTIS? (www.eptis.org)	Is there a PT in EPTIS for these analytes involving this matrix? If so, which one
[21]	Establishment of proficiency testing programs in the Philippines	Infant formula: 38, milk powder: 33, wheat flour: 37, corn-based snack food: 34, powdered concentrate-water-based flavored drinks: 24 These PTs were provided to most of its local and international PT participants for free	Food matrices (infant formula, milk powder, wheat flour, corn-based snack food and powdered concentrate-water-based flavored drinks) for nutrition labeling parameters	Analyses of proximates, minerals, total dietary fiber, saturated fatty acids, cholesterol and vitamin C	No. ISO/IEC 17043 accredited provider	Yes United States (ID: 557098): AOAC International. Related legislation or standard. Germany (ID: 153819): Deutsches Referenzbüro für Ringversuche und Referenzmaterialien GmbH Thailand (ID: 479471): Center for laboratory proficiency testing, Department of Science Service. Minerals in Milk powder United Kingdom (ID: 132354): LGC Standards Proficiency Testing. Dairy Chemistry—(QDCS). United Kingdom (ID: 140970): Fera Science Ltd. FAPAS Food Chemistry. Kenya (ID: 139793): KEBS, Kenya Bureau of Standards. Wheat flour and Maize meal PT scheme Preview PT. Greece (ID: 144477): SCHEMA (Scheme for Chemical Measurement Assessment), General Chemical State Laboratory. nutritional analysis parameters & acrylamide in foodstuffs
[22]	Evaluation of the reproducibility standard deviation in the pesticide multi-residue methods on olive oil from past proficiency tests	A total of 1527 analytical results were collected in the ten PTs	Olive oil	Determination of pesticides in olive oil (QuEChERS method, coupled with LC–MS/MS and GC–MS/MS)	No	Yes Spain (ID: 539076): GSC SL (Gabinete de Servicios para la Calidad). OLIVE OIL 2: Organic and inorganic contaminants [Tested property: Pesticides] Italy (ID: 123559: Lab. Chim. CCIAA Roma. Virgin olive oil "Chemical and organoleptic determinations according to the Reg. CE 2568/91." Tested property: Pesticides

Table 1 (continued)

References	Article title	Number of participants	Matrix	Analytes	Is this PT registered with EPTIS? (www.eptis.org)	Is there a PT in EPTIS for these analytes involving this matrix? If so, which one
[23]	Evaluation of the Use of Microtracers™ in a Proficiency Testing Program	–	Commercially available feedstuffs using base animal feeds and feed additives	Only the recovery of the "Microtracer" was evaluated to verify if it can be applied as a quality control during the division of the prepared samples, thus ensuring good homogeneity	No	No
[24]	First Indonesian proficiency testing using reference value from isotope dilution mass spectrometry method for benzoic acid, methyl paraben, and n-butyl paraben in sweet soy sauce	20	Sweet soy sauce	Benzoic acid, methyl paraben, and n-butyl paraben	No	No
[25]	Proficiency testing as an instrument to assess the analytical performance and the methods routinely implemented: the Italian experience for the screening of antibiotic residues in milk in the official control	31	Milk	Benzylpenicillin (PEN), sulphadiazine (SDZ) and oxytetracycline (OXY)	No	No
[26]	Proficiency testing for total mercury in oyster with a metrologically traceable reference value from isotope dilution mass spectrometry: implications on laboratory practices using mercury analyzers	74	Oyster powder tissue	Mercury	No	No

Table 1 (continued)

References	Article title	Number of participants	Matrix	Analytes	Is this PT registered with EPTIS? (www.eptis.org)	Is there a PT in EPTIS for these analytes involving this matrix? If so, which one
[27]	Provision of proficiency testing for histamine mass fraction in canned tuna to improve the capability of chemical laboratories in the Philippines	12	Tuna paste	Histamine	No	Yes SPAIN—Analysis of histamine in fish https://www.eptis.bam.de/eptis/WebSearch/view/143810; ARGENTINA—Fish determinations—https://www.eptis.bam.de/pts613009; SPAIN—Fishing products—https://www.eptis.bam.de/pts452981; UNITED KINGDOM—Meat and fish scheme—(QMAS)—https://www.eptis.bam.de/pts127280; UNITED KINGDOM—FAPAS Food chemistry—https://www.eptis.bam.de/pts140970
[28]	Results of the 16th proficiency test on the determination of pesticide residues in olive oil	37	Olive oil	27 pesticides (Chlorpyrios, Chlorpyrifos methyl, Lambda-cyhalothrin, Deltamethrin, Diazinon, Diflufenican, Dimethoate, Alpha-endosulfan, Beta-endosulfan, Endosulfan sulfate, Fenitrothion, Fenoxycarb, Fenthion, Fenthion sulfone, Fenthion sulfoxide, Metidathion, Omethoate, Oxyfluorfen, Phosalone, Procymidone, Quinalfos, Kresoxim methyl, Simazine, Terbuthylazine, Tolclofos methyl, Trifloxystrobin, Trifluralin)	No	Yes Virgin olive oil "Chemical and organoleptic determinations according to the Reg. CE 2568/91" https://www.eptis.bam.de/eptis/WebSearch/view/123559
[29]	A proficiency testing scheme to evaluate the effectiveness of laboratory sample reduction of a soil sample	77/76	Soil	Cd, Co, Pb e Mn	No	Yes Cd, Co e Mn (https://www.eptis.bam.de/pts199971)

Table 1 (continued)

References	Article title	Number of participants	Matrix	Analytes	Is this PT registered with EPTIS? (www.eptis.org)	Is there a PT in EPTIS for these analytes involving this matrix? If so, which one
[30]	Acquisition of stability data for pesticides in water sample through proficiency tests	25	Water	109 substances	No	Yes 2,4-D, Dicamba, Mecoprop, Dichlorprop (https://www.eptis.bam.de/pts307812)
[31]	Combining UK and German emissions monitoring proficiency testing data based on stack simulator facilities to determine whether increasingly stringent EU emission limits are enforceable	–	Air	CO, NOx, total organic carbon (TOC), total dust	No	Yes CO, TOC (https://www.eptis.bam.de/pts205029)
[32]	Designing a formulation of synthetic wastewater as a proficiency testing sample: a feasibility study on a laboratory scale	–	Wastewater	$NaHCO_3$, KCl, NaCl and HCl	No	Yes, there are 15
[33]	Fifteen years of proficiency testing of total petrol hydrocarbon determination in soil: a story of success	–	Soil	TPH	Yes https://www.eptis.bam.de/eptis/WebSearch/view/129044	Yes, there are 3
[34]	Measurement of organochlorine pesticides in drinking water: laboratory technical proficiency testing in Mexico	9	Potable water	Aldrin, β-endosulfan, heptachlor, lindane ep and p'-DDE	No	Yes, there are 2

References	Field	Value assignment processes	Variance proficiency/standard uncertainty/normality	Performance assessment	Homogeneity	Stability	Graphical methods
[12]	Food	Consensus value (media)—outliers exclusion with Grubs test	*Z*-score. Intralaboratory precision estimated for all PL using replicate measurements of krill oils in three spectrometers at one laboratory within a short period of time	The raw data were provided by the PT provider. All NMR spectra were additionally evaluated using an in-house developed MATLAB script	It was not done	It was not performed, but samples hydrolyze due to 14 weeks between sample preparation and measurements and different storage conditions related by the authors	PCA of quantitative NMR results of all participants calculated in an automated manner

Table 1 (continued)

References	Field	Value assignment processes	Variance proficiency/ standard uncertainty/ normality	Performance assessment	Homogeneity	Stability	Graphical methods
[13]	Food	The reference activities of ^{137}Cs and ^{40}K were determined by the PT provider. The activity of ^{90}Sr was measured by using liquid scintillation efficiency tracing (CIEMAT/NIST) method in the laboratory of PT provider	Accuracy (U-score) and Precision score	IAEA criteria. The reported results were evaluated against the predefined criteria for accuracy and precision and the status was assigned as "accepted" or "not accepted." Relative bias was calculated as a complementary information and as an additional parameter for the performance evaluation of the participating laboratories	The homogeneity of ^{137}Cs and ^{40}K was tested by analyzing 10 randomly selected plastic bottles, each having about 300 g sample, by taking three independent subsample amounts of 25 g from each bottle and using gamma-ray spectrometry to test the in-bottle and between-bottle homogeneity of the tea powder material; target values of specific activities were determined for each radionuclide. The analysis results were subjected to the Grubbs outlier test, normal distribution, modality test and ANOVA test. A one-way ANOVA approach was considered to evaluate the in-bottle and between-bottle inhomogeneity	Not stated/Not applied	% score reported results and reported results given in the same laboratory order for a complete and easy comparison

Table 1 (continued)

References	Field	Value assignment processes	Variance proficiency/ standard uncertainty/ normality	Performance assessment	Homogeneity	Stability	Graphical methods
[14]	Food	The assigned value was defined by mean value of the participants' analytical results	The results were compared with the obtained with isotopic dilution	The assigned value is suitable to evaluate the analytical performance of each participant relative to all participants. the absolute value, which is not influenced by the analytical methods used by or the skill, etc., of the participants, is required to evaluate the trueness of the participants' analytical methods. As the accuracy of the spiked concentration prepared for the test sample was confirmed in the present study, this value will be a possible solution	The homogeneity of ^{137}Cs and ^{40}K was tested by analyzing 10 samples, taking three independent subsample amounts from each bottle and using gamma-ray spectrometry to test the in-bottle and between-bottle homogeneity; target values of specific activities were determined for each radionuclide. The analysis results were subjected to the Grubbs outlier test, normal distribution, modality test and ANOVA test. A one-way ANOVA approach was considered to evaluate the in-bottle and between-bottle inhomogeneity	Before and after the experiment (PT)	Not stated/Not applied
[15]	Environmental	Spectrometry with HPGe-detectors. TAEA did a direct comparison of peak areas, corrected for density and composition, with reference material clover sample (IAEA-156)	Not stated	Evaluation of all PT rounds use the absolute value of the z-score and the best result were 77% reported the result of samples had an absolute z-score lower than 2	Not stated/Not applied	Not stated/Not applied	*z*-score graphs for individual sample evaluation and Performance evaluation (by percentage of satisfactory results)
[16]	Environmental	(0.605 ± 0.024) Bq/g ^{40}K, (1.003 ± 0.015) Bq/g ^{226}Ra and (0.806 ± 0.013) Bq/g ^{228}Ra	There is the calculation of material uncertainty (but the origin of the calculation is not displayed)	Not evaluated	The homogeneity test was not performed, considering the samples were homogeneous. When the samples deviating too much from the mean, they were excluded	Cursory sample stability measurements taken over the course of a few weeks suggested qualitatively that the material is sufficiently stable as PT sample	It gives the outlier-corrected results of the proficiency test in relation to BEV's reference value

Table 1 (continued)

References	Field	Value assignment processes	Variance proficiency/ standard uncertainty/ normality	Performance assessment	Homogeneity	Stability	Graphical methods
[17]	Environmental	Comparison of the laboratory-reported value against the assigned values using the percentage difference and zeta tests (ISO 13528:2005, Harms, A.V., 2009);	Determination of uncertainty ratio outliers using the interquartile range tests	Grading of laboratory performance using the grading scheme: Accuracy, Zeta-score; Correct identification of nuclides, Precision	Not stated	Not stated	Not stated
[18]	Environmental	The consensus value was obtained by the robust mean of the logarithmic values following standard ISO 13528 including a robust standard uncertainty	The relative uncertainty of the assigned value per component was determined by considering the aforementioned robust relative standard uncertainties for the consensus threshold values of each component as well as the relative uncertainty of the dosed mass concentration	The evaluation of the participants' results was done on the basis of a z-score procedure after logarithmic transformation	Not stated	Not stated	Recovery rates for each component per dosage, participant, and proficiency test Mean *z*-scores per component, participant, and proficiency test (PT) Comparison of the z-scores of the other components with *n*-butanol together with angle bisector, and correlation coefficient r

Table 1 (continued)

References	Field	Value assignment processes	Variance proficiency/ standard uncertainty/ normality	Performance assessment	Homogeneity	Stability	Graphical methods
[19]	Food	The assigned values were established independently of the results reported by participants as they were derived from the gravimetric preparation	The associated standard uncertainties of the assigned values ($u(x_{pt})$) were calculated following the law of uncertainty propagation, combining the standard measurement uncertainty of the characterization (uchar) with the standard uncertainty contributions from homogeneity (uhom) and stability (ust) studies (Table 1), in compliance with ISO 13528:2015 (ISO 13528, 2015). The standard deviation for proficiency assessment, σpt (Table 1), was set using a maximum tolerated standard uncertainty of 25 % following the Guide 11813/2017	The individual laboratory performance was expressed in terms of *z*-scores according to ISO 13528:2015 (ISO 13528, 2015)	The homogeneity experiment consisted of duplicate analysis on 10 samples randomly selected along the filling sequence, according to ISO 13528:2015. The test material was rated adequately homogenous at a sample intake of 5 g and no trend was observed	The stability of the test material was evaluated following the requirements in ISO 13528:2015. Nine randomly selected samples were stored at different conditions for 3 weeks, covering the whole period of the PT exercise, from the dispatch of the PT items to the end of the submission of the results. No significant differences in the analyte contents of the test samples were found. Hence, adequate stability of the test samples over the whole period of the study can be assumed, provided that the recommended storage conditions were applied	Graph with measurement results and associated expanded uncertainty (as reported by participants, Sample A). Assigned range (xpt ± *U* (xpt); acceptance range (xpt ± 2σpt). Youden confidence ellipse graph. The ellipse shows the bivariate confidence limit at the 95% confidence level. The numbers indicate laboratory codes
[20]	Food	Not stated, Not applicable	Comparison between relative standard deviation (RSD) of proficiency tests and collaborative tests	Not stated, Not applicable	Not stated, Not applicable	Not stated, Not applicable	Linear regression plots between the standard deviation of reproducibility (RS) versus mass fraction (w) for PT and CT results for the analyzed analytes

Table 1 (continued)

References	Field	Value assignment processes	Variance proficiency/ standard uncertainty/ normality	Performance assessment	Homogeneity	Stability	Graphical methods
[21]	Food	The consensus value was derived from the robust average (x^*) of PT participants' results computed using Algorithm A of ISO 13528 standard	Uncertainties ($u(x_{pt})$) and the standard deviation for proficiency assessment (σ_{pt}) were derived for the results of participants of the corresponding PT rounds. Algorithm A of ISO 13528	The z'-score was used when $u(x_{pt}) \geq 0.3\sigma_{pt}$ as suggested in the ISO 13528:2015	The statistical evaluation of homogeneity studies was performed according to ISO 13528:2015 and Fearn and Thompson	Three randomly selected proficiency PT items, stored at appropriate temperature conditions for a period of time (i.e., 1.5 and 3 months for infant formula, wheat flour and corn-based snack food; 1.5 and 4 months for milk powder; and 2 months for powdered concentrate), were analyzed in duplicate for all relevant measurands by the same subcontracted laboratories, using the same experimental conditions applied in the frame of the homogeneity study	Did not use graphical methods; only tables with summary statistics and properties of the proficiency test item, per round of PT
[22]	Food	The assigned values (xpt) were obtained using robust statistics, namely, logarithm A described in ISO 13528: 2015	Standard uncertainties (u (x_{pt})) were obtained by applying robust statistics, namely, logarithm A described in ISO 13528: 2015. The standard uncertainty was calculated using the robust standard deviation (s^*) and the total number of results (n)	Not stated, Not applicable	Not stated, Not applicable	Not stated, Not applicable	Graph for the evaluation of the robust relative standard deviation (RRSD,%) in the last ten years graph with the trend of the ratio between the uncertainty of the reference value and the assigned standard deviation used in the calculations of the z-scores graph for assessing veracity (via% recovery) versus concentration levels (mg/kg)
[23]	Food	Not stated	Not stated	Not stated	Not stated	Not stated	Not stated

Table 1 (continued)

References	Field	Value assignment processes	Variance proficiency/ standard uncertainty/ normality	Performance assessment	Homogeneity	Stability	Graphical methods
[24]	Food	Primary method (validated) to obtain the reference value (GC/MS—isotope dilution method). Replicate analysis of 10 samples chosen at random using single-point calibration	Proficiency standard deviation was obtained by the Horwitz equation. The standard uncertainty was calculated according to ISO Guide 35, considering homogeneity, stability and characterization. The uncertainty of the robust average was also calculated as informative, according to Algorithm A of ISO13528	*Z*-score if the criterion "standard uncertainty of the assigned value < 0.3 standard deviation for proficiency assessment" is met and z'-score if this criterion is not met	Ten samples were evaluated analyzed under repeatability conditions. The criterion for the homogeneity check was "between-sample standard deviation ≤ 0.3 standard deviation for proficiency assessment" (in accordance with ISO 13528)	Isochronous method for 4 weeks (1 sample taken each week)—Storage at 4 °C and analysis at 25 °C. Stability during transport: storage at 40 °C before sending samples. Criterion: "laverage of the homogeneity check results—average of the stability check results| ≤ 0.3 standard deviations for proficiency assessment."	Histogram for *z*-score. Results with the uncertainty bars in a graph with the concentration range for approval
[25]	Food	The assigned values were calculated from the weighing and dilution steps applied to prepare the materials	Not stated	A score ranging from 0 to 1. A score of 1 was assigned for four correct final results (with the total of five). This score was decreased by 0.25 for each false negative (FN) result and by 0.125 for each false-positive (FP) result, in agreement with EURL for antimicrobial residues in food from animal origin scoring criteria for qualitative PTs. 1.00 (satisfactory), 0.75 (questionable), 0.50 (unsatisfactory)	Not stated	Not stated	Not stated

Table 1 (continued)

References	Field	Value assignment processes	Variance proficiency/ standard uncertainty/ normality	Performance assessment	Homogeneity	Stability	Graphical methods
[26]	Food	The mass fraction of mercury in the PE sample was within Korean law (0.25 mg/kg). The resulting reference mass fraction and its associated expanded uncertainty were (0.246 ± 0.012) mg/kg ($k = 1.96$, with a 95% confidence level)	The Horwitz equation was used to derive the standard deviation for proficiency assessment	Through the score of each reported mass fraction. Robust analyses were performed using the ISO 13528 Algorithm A	Twelve sample bottles were systematically selected at regular intervals to assess homogeneity	Storage conditions: the mercury measurement was repeated with three bottles of oyster powder material 1, 3, 6 and 12 months after certification. Stored at temperature: 24 °C ± 2 °C, humidity: 55 % ± 20 %. Three solutions were prepared at each concentration level	Scatter plots similar to the control cards with z-score plotted in the ordinates and the participants in the abscissa; box diagram (or also called "action graphics" by excel)
[27]	Food	Two batches of test materials were produced containing two different levels of histamine concentration in canned fish with the assigned values of 148 and 65 mg/kg for the EP of the years 2014 and 2015, respectively. Homogeneity and stability tests (short and long term) were performed using two HPLC methods with a fluorescence detector (340 nm for excitation and 445 nm for emission)	The uncertainty of the assigned value was calculated according to the GUM (> 41% for characterization only). The reported expanded uncertainties of the reference values were estimated using the ISO GUM approach. The combined standard uncertainty (u) of the reference value was evaluated from uncertainties obtained in the characterization, homogeneity (u_{bb}) and stability (u_{lts}). The expanded uncertainty (U) was calculated considering $k = 2$	By means of score $z = (xi—xpt)/\sigma pt$, where xpt is the assigned value and σpt is the standard deviation for proficiency assessment, derived from the Horwitz equation	For each batch (total 2 with different histamine levels) 10 vials were selected for the study of homogeneity. The statistical analysis of the homogeneity study was carried out according to the Harmonized Protocol of IUPAC (visual inspection, Cochran's test at 95 % confidence level and analytical variance estimate—ANOVA). The ratio between analytical precision (s_{an}) and standard deviation for proficiency assessment (s_{an}/σ_{pt}) was found below the prescribed value of 0.5. Comparison between the variation between samples and the critical test value	Short term: isochronous approach for 3 weeks at 40 °C, 30 °C and 4 °C. The investigated storage periods were 1, 2 and 3 weeks, using 2 samples for exposure to temperature. Samples stored at -20 °C were used as control samples. Five replicates of all samples exposed to different conditions and periods of exposure were analyzed by HPLC. Long term: investigated at 4 °C, they were performed for up to 12 months. Trend analysis described in ISO Guide 35 was used to assess the stability of the two materials	Histogram for uncertainty and z-score

Table 1 (continued)

References	Field	Value assignment processes	Variance proficiency/ standard uncertainty/ normality	Performance assessment	Homogeneity	Stability	Graphical methods
[28]	Food	The absence of the 27 pesticides was verified in an olive oil sample (it does not mention by which method) and the samples were doped in the mass fraction range with the doping (from 0.050 mg/kg to 0.350 mg/kg)	All calculations were performed according to ISO 13528: 2015. To find the uncertainty of the assigned value, robust averages (x_{pt}) and robust standard deviations (s *) were calculated using Algorithm A of the same standard	The individual performance of the laboratory was expressed in terms of z-score according to ISO 13528: 2015. The z-score was calculated using the robust means and the algorithm. No z-score was calculated for false-positive results. In addition, an overall performance was assessed by calculating the mean of the z-squared scores	10 random flasks were analyzed in duplicate. Statistical evaluation according to ISO 13528 proved homogeneity	Stability was proved by the classic method with two bottles (randomly chosen), which were analyzed in duplicate on two occasions: before the samples were shipped (day 1) and just after the deadline for disclosing the results (day 2). There was no significant decrease in the investigated pesticides	Histograms for the mass fraction of chlorpyrifos and endosulfan sulfate. Histogram for the overall performance of laboratories, calculated by averaging the squared z-score. Histograms for comparing the z-scores for the results of alpha-endosulfan and beta-endosulfan of the current EP with the previous one
[29]	Environmental	Tests t and F were performed with a 95% confidence level (in all cases the averages and variance showed significantly different results)	The substantial average and standard deviation were estimated according to ISO 13528	Through the z-score (Cd, Co and Pb showed negative results and Mn showed a positive result)	Regression analysis of the reported result opposite to the bottling order failed to identify a concern with homogeneity between samples	Not stated	Box plot for each metal
[30]	Environmental	Not stated	Not stated	For assigned values and z-scores, comparisons between the results at the different dates of analysis were made	Not stated	The variation coefficients of the study showed the reliability, which can be referred to the assigned values to evaluate the stability of the substance	Linear regression graph for the result of participants' stability
[31]	Environmental	Not stated	Associated expanded uncertainties required for CO, NOx, TOC and total dust in accordance with, EN 15058, EN 14792, IED and EN 13284–1:2015, respectively	The z-scores of each matrix were evaluated according to their legislation (CO: EN 15058:2006; NOx: EN 14792:2005; TOC: EN 12619:1999; Total Dust: EN 13284–1:2017)	Not stated	Not stated	Box plot for evaluation of Z-score

Table 1 (continued)

References	Field	Value assignment processes	Variance proficiency/ standard uncertainty/ normality	Performance assessment	Homogeneity	Stability	Graphical methods
[32]	Environmental	Conductivity value: 6.67 mS cm^{-1} and uncertainty of 0.03 mS cm^{-1} (0.46%) at 25 °C	The uncertainty of TDS, EC and the conversion factor are calculated based on the standard deviation of five measurement repetitions	No	Tests were conducted based on ISO 13528: 2015. The sample must meet the evaluation criteria where s (sample) $\leq 0.3\sigma_{pt}$ and unidirectional ANOVA with 95 % confidence level, where σ_{pt} is the standard deviation for proficiency assessment	Tests were conducted based on ISO 13528: 2015. The material candidate PT is considered to be adequately stable. And if Y_1-Y_2 What other people are saying $\leq 0.3\sigma_{pt}$, where Y_1 and Y_2 are the average conductivity measured in the first and the fourth week, respectively	No
[33]	Environmental	The average *X* was taken as the assigned Xref value for the proficiency assessment together with the standard deviation for proficiency assessment, according to ISO 13528	Not stated	*Q* score (percentage difference)	Statistical evaluation, according to ISO 13528 proved homogeneity	Not stated	Dispersion graph, *Y* axis with *Q* score and X axis with the years
[34]	Environmental	Calculated from the reference values determined by the organizer, analyzing six samples of fortified water according to the procedure described above and extraction of the analytes according to the EPA 508 method	Standard deviations appropriate for proficiency, per analyte, were obtained from the Horwitz equation	*z*-score: The acceptable *z*-scores obtained by the participating laboratories were identified	The unidirectional analysis of variance (ANOVA), as recommended in the ISO Guide 35 Guide, *F* Values	The stability tests were considered acceptable if the difference in concentration for each pesticide was less than 10%	Yes, there is the use of bar graphs with *Y* axis containing *z*-scores and *X* axis containing participating laboratories

1. Poor statistical analysis of the data;
2. Original experiments performed with low repeatability;
3. Poor experimental design; and
4. Raw data unavailable from the original laboratory.

The adoption of the general requirements according to ISO/IEC 17025 forces the application of quality tools by testing laboratories, such as calibration, CRM application, validation of methods, and measurement uncertainty estimation, which can significantly improve the reproducibility of the published experiments. However, only interlaboratory comparisons (such as periodic participation in PT) can assess the performance of laboratories and the agreement among them [4], showing the capacity of the laboratory to reproduce its results.

Critical steps of PT planning and development

The planning to perform PT is based upon several stages, carried out in a logical sequence to achieve the specifications of each program. Among these stages, it is important initially to establish how the material will be produced. As an example, one can choose a naturally contaminated or spiked sample. Different techniques can be used to calculate the assigned value, such as by a primary method in the laboratory, two different techniques, or a consensual value between results reached from PT after adequate statistical evaluation. The homogeneity and stability must be evaluated to ensure comparability of all results. Based on the results of participants, the evaluation needs to consider issues such as the normal distribution and standard deviation of data, and graphic evaluation of results.

Considering ISO 13528, the provider of PT has to ensure that all PT items are stable and homogenous. However, the guide to express the uncertainty in measurements (GUM) shows that uncertainty calculation implementation is required to assure adequate stability and homogeneity. Due to lack of repeatability in measures or insufficient repetitions, some studies of homogeneity and stability fail to properly interpret the results [38]. The standards ISO/IEC 17043 and the ISO Guide 35 [39] demand uncertainty calculations linked to homogeneity, stability, and characterization to reach a standard uncertainty for PT.

During evaluation of PT, graphical methods also help interpret the testing results, enabling laboratories to implement corrective actions. [40].

The generalized need for PT programs to be accredited brought the need for variations in the evaluation of z-scores to encompass all statistics of the tested PT items. Thus, it is essential to consider other parameters, described by the ISO 13528 standard as z'-scores, Zeta-scores, D-scores, and En scores. However, in most cases z-scores are sufficient to evaluate the performance by a PT provider [41].

Considering the different statistical tools applied in different PT steps, we evaluated 23 papers published in the last two years (2018–2019), focusing on food and environmental matrices due to their importance and representativeness, as discussed above. The results are summarized in Table 1 and are discussed below.

Homogeneity

According to requirement 6.1 of ISO 13528-2015, "The PT provider shall ensure that batches of PT items are sufficiently homogeneous and stable for the PT testing scheme. The provider shall assess homogeneity and stability using criteria that ensure that inhomogeneity and instability of PT items do not adversely affect the evaluation of performance."

The assessment of homogeneity and stability should use different approaches, as detailed in ISO. Apprising the manuscripts chosen in this review, we observed different ways to assess the homogeneity and stability. Therefore, we present a discussion of these descriptions in the next sections.

PT providers have different ways to ensure the homogeneity of PT items subjected to testing. Some of them do not consider homogeneity, such as PTs of krill oil samples [12] and screening of antibiotic residues in milk [25], due to the samples characteristics. The same was observed mainly in PT schemes focusing on radionuclide samples. Nakashima and Duran [17] requested a contractor to verify the uniform distribution of radioactivity in a filter. The uncertainty due to homogeneity during method validation should be included and disclosed. Other authors homogenized analytes manually after spiking test samples [14]. Even though not performing a homogeneity test, gamma-spectrometric measurements in predefined lines were performed to check for homogeneity issues in the prepared quartz sand for different grain size fractions. The homogeneity of the sample was studied only between containers, not within containers. The final uncertainties of the respective sums of relative deviations were calculated using Gaussian propagation of uncertainties [16].

Kim et al. [26] evaluated the homogeneity of mercury in samples of oysters by isotope dilution mass spectrometry. Twelve bottles of PT items were selected, and the isotope ratios were used, ^{202}Hg and ^{200}Hg, estimated based on the intensity ratios of the two isotopes ($I(^{202}\mathrm{Hg})/I(^{200}\mathrm{Hg})$) by ICP-MS. The isotope-enriched material and a standard solution with a certified mass fraction were used in the model equation to estimate the measurand. The laboratories' development of adequate homogenization of samples under different conditions (good between-sample but poor within-sample homogeneity) was evaluated in a PT to measure metal contents in soil samples [29].

Yeltepe et al. [13] tested the within-bottle and between-bottle homogeneity by gamma-ray spectrometry. Target

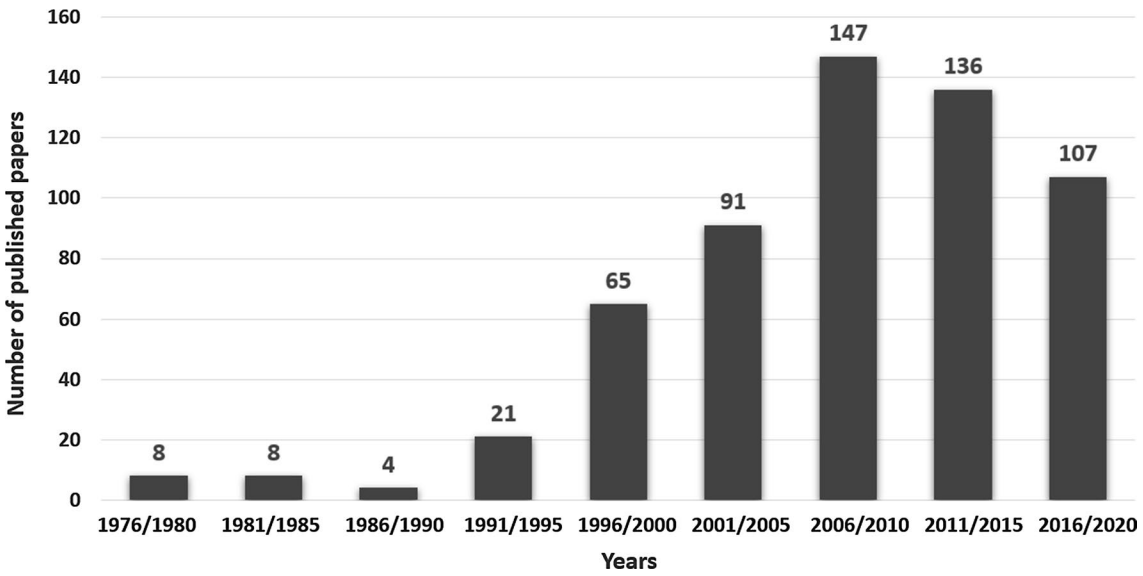

Fig. 4 Frequency of publications with the expression "proficiency testing" contained in the title considering the Web of Science platform for the major areas of chemistry, physics and biology between 1977 and 2020

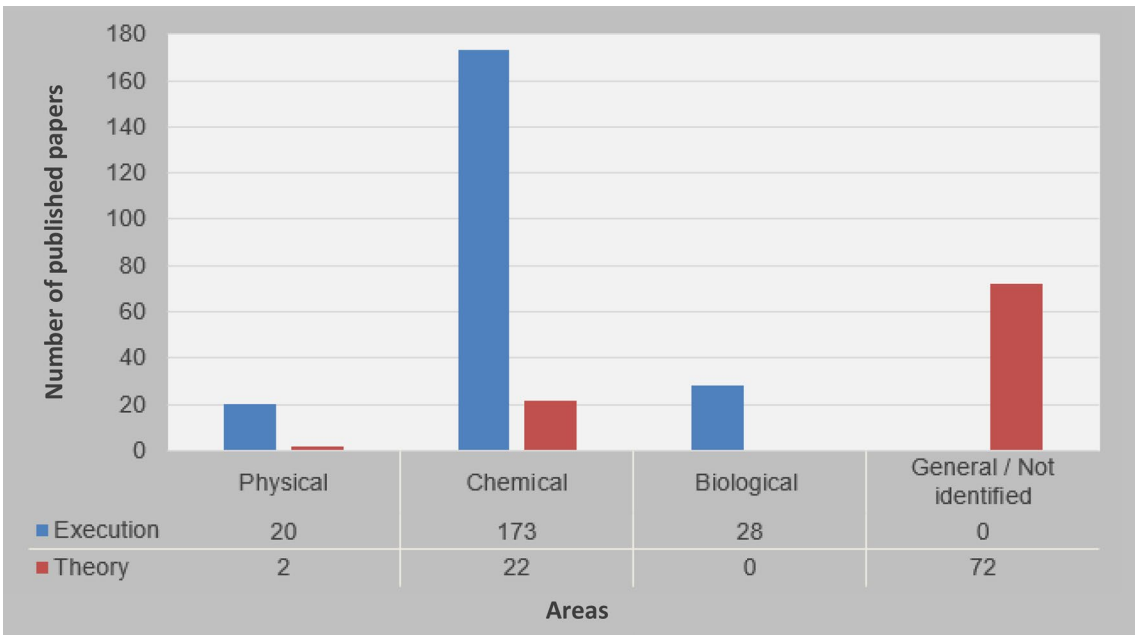

Fig. 5 World trends in publications about PT (period from 2009 to 2019 in the three databases: Web of Science, ScienceDirect, and Scopus)

values of specific activities were determined for each radionuclide. One-way ANOVA was applied to evaluate the in-bottle and between-bottle homogeneity [14]. Another statistical procedure for homogeneity evaluation is to apply the IUPAC Harmonized Protocol [40]. Visual inspection, the Cochran's test, and an estimate of analytical variance (ANOVA) have also been used to test whether samples' homogeneity was achieved [27].

Among the PT studies reporting the use of ISO 13528, Cordeiro et al. [19] evaluated official control laboratories' performance in quantifying fipronil reliably in eggs. The homogeneity experiment consisted of duplicate analysis of 10 PT items randomly selected along the filling sequence, and the analyses were performed in random order. The same procedure was employed to evaluate pesticide residues in olive oil [28], total petrol hydrocarbons in soil [33].

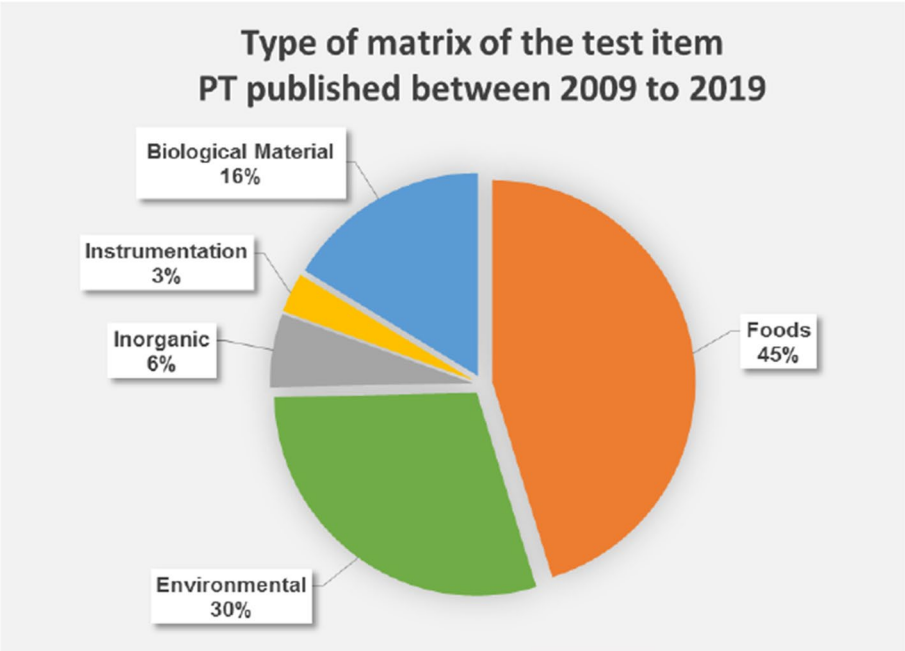

Fig. 6 Percentage distribution of PT papers classified as "execution" considering different matrices used

Furthermore, PT was applied using isotopic dilution to measure benzoic acid, methylparaben, and *n*-butyl paraben in sweet soy sauce [24] and to measure electrolytic conductivity in a synthetic formulated wastewater for water quality monitoring [32].

Following ISO 13528, the criterion for the homogeneity check was "between-sample standard deviation ≤ 0.3 for proficiency assessment (s(sample) $\leq 0.3\sigma_{PT}$)" and one-way ANOVA at 95% confidence level. Fearn and Thompson [42] proposed an alternative statistical approach to correct errors in the test suggested in the Harmonized Protocol. They argued that some materials that are rejected are in fact satisfactory. To overcome this drawback, the authors presented a simple new experiment, namely a randomized replicated experiment using ANOVA to estimate the sampling standard deviation (σsam) from the results. In the experiment, the sample was homogenized and analyzed in duplicate. If the analytical method is precise, the standard deviation (σsan) is small. This way σsam can be reliably estimated. The authors suggested that the analytical precision of the homogeneity test method should satisfy $\sigma_{sam}/\sigma_{pt} < 0.5$.

ISO Guide 35 was applied to evaluate the homogeneity in a PT conducted to evaluate detection of organochlorine pesticides in drinking water. One-way analysis of variance (ANOVA) was used to evaluate water samples' homogeneity concerning the concentration of pesticides [34]. It is necessary to emphasize the similarity between the established homogeneity and stability tests in the ISO Guide 35 and ISO 13528. Otherwise, some points can present differences. ISO Guide 35 includes both within- and between-unit homogeneity. Statistical power analysis can assist in choosing a suitable number of units and replicates for the homogeneity study. Also, in ISO Guide 35 the uncertainty should be calculated because any detected heterogeneity is included in the certified value. On the other hand, ISO 13528:2015 is less restrictive. Usually, only two replicates are performed. A comparison between them is made, and if necessary, a statistical test (e.g., Cochran's test) is applied for outlier exclusion. The between-sample standard deviation, S_s, is compared with the standard deviation for proficiency assessment σ_{pt}, and the PT items can be deemed adequately homogeneous if $S_s \leq 0.3\ \sigma_{pt}$.

The justification for the factor of 0.3 is that when this criterion is met, the between-sample standard deviation contributes less than 10 % of the variance in evaluating the performance. Hence, the performance evaluation is unlikely to be affected.

Other authors did not refer to ISO standards. Instead, the use of CRMs for comparison in INAA and k0-INAA experiments using PT materials is one of the alternatives to the homogeneity test described in ISO 13528 [15].

Stability tests

As described in requirement 6.1.2 of ISO 13528, "for calibration proficiency testing schemes where multiple participants use the same artifact, the proficiency testing provider shall assure stability throughout the round, or have procedures to identify and account for instability through the progression of a round of the proficiency testing scheme. This should include consideration of tendencies for particular PT items and measurands, such as drift. Where appropriate, the

assurance of stability should consider the effects of multiple shipments of the same artifact."

Concerning the evaluated revised articles, we observed different options related to the stability process. Zailer et al. [12], in a PT to evaluate ^{31}P NMR as the official method for measurement of phospholipid in krill oil, did not perform a stability test. The authors reported problems related to the hydrolysis process, in which PtdCho was gradually transformed into degradation products.

As proposed by ISO 13528, stability was evaluated in a three-year trial of isotope dilution mass spectrometric results in a PT for pesticide residue analysis [14].

Wiedner et al. [16] performed cursory sample stability measurements for a few weeks to suggest a prepared technologically enhanced naturally occurring radioactive material ((TE)NORM), used in a PT of gamma-ray spectrometry laboratories, finding it to be sufficiently stable as PT sample.

Two main strategies have been applied to evaluate stability. One of them is the isocratic experiments, where the samples are stored in different temperatures and humidity in laboratory conditions, and all analyses are performed simultaneously. This scheme was applied based on ISO 13528 to assess pesticide stability in eggs [19] and based on ISO Guide 35 to evaluate the stability of two reference materials of canned tuna containing histamine. In this issue, short- and long-term stabilities were studied. Short-term stability was satisfied at the three evaluated temperatures, and the materials were considered stable during the transport. This result was obtained as $b_1 < t_{(0.95,\, n-2)}$, where b_1 is the regression line's slope, $t0{:}95$, and $n - 2$ is the Student's t value at 95% confidence level with $n - 2$ degrees of freedom. A similar experiment was performed for the long-term stability at 4 °C, with the histamine determined after 12 months. The results were included in the combined standard uncertainty (u) of the designed reference values [27].

Stability has also been determined in PT studies with analyses before the PT items are distributed to laboratories and at different periods. Some PT providers analyze the PT items in the middle and at the end of the test period, such as a 16-year study to evaluate the stability of six different pesticides added in a commercial olive oil. Two bottles were randomly chosen and analyzed before shipping to a PT and in the day defined for reporting results [28]. However, most stability studies evaluate samples at different times and considering different temperatures. Samples of oyster powder and solutions were stored in low-density polyethylene (LDPE) bottle containers to check the stability of Hg during PT. Mercury was determined in 1, 3, 6, and 12 months after certification [26]. Robust means (Algorithm A of ISO 13528) was applied to assess the stability of more than 100 pesticides in freshwater. The samples were analyzed 2, 4, and 9 days after preparation, and most of the pesticides were found to be stable during the considered test period [30].

Difference between measures $\bar{y}1 - \bar{y}2 \leq 0.3_{\sigma PT}$, where $\bar{y}1$ and $\bar{y}2$ are the averages of measurement and σ_{PT} is the standard deviation of the PT, has also been applied to assess the stability. This proposal was used to determine the best formulation for a PT candidate material of synthetic wastewater. The samples were stored at two different temperatures (25 °C and 40 °C), and the conductivity was analyzed in the first and fourth weeks [10, 32]. The same strategy was used to determine nutrients stability in five different foods exported by the Philippines. Three randomly chosen samples were separated, stored at controlled temperature, and analyzed after 1.5, 2, and 3 months. The stability considered the $|\bar{y}1 - \bar{y}\,2|$ for each measurement [21].

Assignment of target value—x_{pt}

The PT target value is considered the best estimate of the true value, and it is one of the most critical steps defined by the provider. Sections 7.3 and 7.7 of ISO 13528 specify statistical procedures and models to determine the target value, but the choice of these procedures is the PT provider's responsibility. Nevertheless, their documentation and disclosure to participants are mandatory.

The sample formulation through the standard addition was one of the strategies utilized in the PT items, and the target value defined by the provider was the average of the PT results after the exclusion of outliers through the Grubbs test [12, 34] or the robust average (x*) calculated from Algorithm A (ISO 13528—item C.3.1) [22, 28, 29].

Also considering the PT items' preparation through standard addition, Yarita et al. [14] and Ebarvia et al. [27] characterized the materials with the application of the primary method through isotope dilution, which is a procedure used when the target value is defined by only one laboratory (ISO 13528—item 7.3.1.4).

We also observed that some of the providers utilized reference materials (RMs) as PT items. Wiedner et al. [16] produced a reference material (naturally occurring radioactive materials—NORM) constituted of quartz sand with high levels of ^{226}Ra, which was characterized through gamma-ray spectrometry as a primary method by the National Metrology Institute of Austria, and this reference value was compared to the participants' average.

Kim et al. [26] produced a CRM through standard addition, and the characterization was realized using the primary method of isotopic dilution, following the predicted procedures of ISO Guide 35. This CRM value was employed as a target value compared to the robust average obtained through Algorithm A (ISO 13528—item C.3.1).

Ziegler et al. [30] spiked water samples to evaluate pesticide stability through a PT, observing that 81 % of evaluated analytes were within 20 % of the theoretical spiked value.

Concerning the procedure that defines the target value from the consensus value of the PT participants, most of the studies mention Algorithm A (ISO 13528—item C.3.1) to obtain the robust average (x^*) as a target value [22, 28, 29]. Becker et al. [33] utilized the Hampel robust estimator to determine the target value, as specified by ISO 13528—item C.5.3.

Based on the Wageningen Evaluating Programs for Analytical Laboratories (WEPAL) PT results, other authors evaluated instrumental neutron activation analysis (INAA) as the primary method. Its performance was evaluated by comparing the values reported by the laboratories, using the average of all laboratories as the target value, considering a regular distribution [15]. Coleman et al. [31] used other PT items and adopted the pre-established target value by those providers.

Calculating standard deviation of PT evaluation (σ_{pt})

In general, the standard deviation of PT evaluation must delimitate the confidence interval as the difference between the value reported by the laboratory and the target value, assigning a score in a specific confidence interval.

In the same way as the target value, the standard deviation of the PT evaluation can be defined using a statistical model suggested in requirement 8 of ISO 13528. It is important to define whether the σ_{pt} value will be calculated from the results reported by the laboratories or defined independently.

Among the procedures that utilize the classical statistic, Ziegler et al. [30] defined the σ_{pt} value from the estimate of the standard deviation of the integrated signals of the nuclear magnetic resonance. In this study, each sample spectrum was submitted to the Grubbs test, at $\alpha = 0.05$ significance level for outlier exclusion. The target value and its standard deviation were used as the tolerance limit to evaluate each laboratory's result with its standard deviation [16].

As described in requirement 6.5.1 of ISO 13528, robust methods are recommended to PT providers to define the standard deviation value, preferably through methods that exclude outliers. Thus, we observed in this review that some authors employed the normalized interquartile range (ISO 13528—item C.2.3), Algorithm A (ISO 13528—item C.3), and the Hampel estimator (ISO 13528—item C.5) as robust statistical techniques to estimate the PT standard deviation value.

Stöckel et al. [18] determined the standard deviation of PT calculated from the Hampel robust estimate. Since this procedure reduced the impact of outliers on the results, the authors found better estimates for the reproducibility standard deviation, in which one of the goals was the evaluation of this merit parameter. Another robust test used was the normalized interquartile range, applied to exclude outliers in the uncertainty calculation of laboratories' performance evaluation through the Zeta-score [17].

Currently, Algorithm A is widely used for the PT standard deviation calculation because it is based on descriptive statistics. The average and the standard deviation are calculated interactively, replacing the outlying values several times. Algorithm A transforms the original data through a process called winsorization to provide alternative average and standard deviation estimators for almost regular data, which is more useful when the expected proportion of outliers is below 20 %. The winsorization process is based on replacing the extreme values by the maximum and/or minimum values in the dataset [43]. The convergence of the average and standard deviation values can be assumed when there is no change of an interaction in the third significant number of the robust average and the robust standard deviation (ISO 13528—item C.3).

Generali et al. [28] and Stefanelli et al. [23] calculated the PT standard deviation (σ_{pt}) with the results derived from laboratories that participated in a testing round. The authors utilized Algorithm A to estimate the robust standard deviation (s^*) until the interaction stabilized at the third significant figures. The robust standard deviation was defined as the PT standard deviation, which was used for the participant laboratories' performance evaluation.

As specified in ISO 13528—requirement 8.4, other authors used the Horwitz equation, which establishes an exponential relation between the standard deviation of reproducibility and the mass fraction of analytes. This is an empirical parameter applied to reproducibility standard deviation evaluation, considered a general model to designate the standard deviation of the PT evaluation [44]. This equation is utilized to predict the chemical measures of interlaboratory tests, and it was proposed by Horwitz [45] and modified by Thompson et al. [40].

Leyva-Morales et al. [34] and Kim et al. [26] used the Horwitz model to calculate the PT standard deviation, which was determined from the spiked sample's target values, considering the analytes of interest. Kim et al. [26] found a high level of equivalence between the values reported by the participant laboratories and the reference value (target value), assigning the PT standard value from the Horwitz predictive model, which was satisfactory to obtain conclusions about the potential of the TDA-ASS method for measurement of mercury levels.

Likewise, Ebarvia et al. [27] assigned the standard deviation calculated from the Horwitz model. For the intended purpose, it was possible to observe a significant improvement in the evaluated laboratory performance. This confirmed that the Horwitz model is a good option

for PT providers to assign the standard deviation of the PT evaluation.

Standard uncertainty calculation of the target value ($u(x_{pt})$)

Generali et al. [28] and Stefanelli et al. [22] calculated the target standard uncertainty value ($u(x_{pt})$) according to the equation: $u(x_{pt}) = 1.25 \times \frac{s^*}{\sqrt{p}}$, where p is the number of results and S^* is the robust standard deviation. The authors determined that $u(x_{pt})$ was below the criterion $u(x_{pt}) < 0.3\sigma_{pt}$, so it was considered insignificant. Therefore, it did not need to be included in interpreting the results of the PT round. However, according to the established criterion, Dajay et al. [21] found nonsignificant standard uncertainty values of the target value for protein, ash, and sodium analytes, as specified in ISO 13528—item 9.2. But for moisture as well as the analytes fat, iron, calcium, potassium, and zinc, the study presented significant uncertainties—$u(x_{pt}) > 0{,}3\sigma_{pt}$, which were assigned in the laboratories' performance evaluation.

Applying a different approach, instead of considering the PT results, Aryana et al. [24] assigned the target standard uncertainty value from the combination of PT items with uncertain characterization (u^2_{char}), the homogeneity test standard uncertainty (u^2_{hom}), and the standard uncertainty due to instability (u^2_{stab}), according to the formula:

$$u(x_{pt}) = \sqrt{u^2_{char} + u^2_{hom} + u^2_{stab}}.$$

Performance evaluation

The statistical project to be applied by PT providers for performance evaluation must provide objective information that shows the quality of the participant laboratories' results based on the criteria pre-established through interlaboratory comparison.

ISO 13528 requirement 8 proposes PT evaluation criteria that generally must be established to meet the following purposes: performance evaluation through comparison with external criteria; performance evaluation through interlaboratory comparison; and performance evaluation through comparison with the stated measurement uncertainty.

One of the procedures described in ISO 13528 compares each participant's reported results to a target value, establishing a confidence interval through the PT standard deviation. In this case, the standardized performance statistic usually is the z-score or z'-score, which allow direct comparison of the results of the different PT items and different units, since the score is not expressed in the measurand's original unit, that is, it is normalized and described as the distance between the participant laboratory's results and the target value in standard deviation units. The hypothesis of using the z-score or z'-score is based on the dataset's normal distribution, with averages of 0 (zero) and 1 (one) standard deviation.

Among the evaluated works, the z-score was used for environmental matrices [15, 18, 31, 34] and food matrices for the purpose of performance evaluation [19, 26–28, 30]. The z'-score was cited by Dajay et al. [21] when the target value standard uncertainty ($u(x_{pt})$) was greater than $0.3\sigma_{pt}$, according to the criterion established in ISO 13528.

Leyva-Morales et al. [34] identified the need to harmonize methods for analysis of the levels of organochlorine pesticides in potable water; Stöckel et al. [18] reported that the participants' performance evaluation based on the z-score after logarithmic transformation of the results was satisfactory.

Generali et al. [28] concluded that the combined z-score must be used with caution and that the ideal is use of the individual z-score for performance evaluation.

The Zeta-score (ζ) is also used as a statistical criterion [13, 17, 24]. The Zeta-score is mainly applied when the PT goal is to evaluate laboratories' performance in comparison with a target value associated with an assigned uncertainty. The performance criterion is interpreted in the same way as the z-score (ISO 13528 requirement 9.6). Aryana et al. [24] evaluated laboratories' performance in detecting benzoic acid, methylparaben, and n-butylparaben, and the results (60 %, 69 %, and 54 %, respectively) were considered satisfactory.

Other performance estimators, like the normalized error (E_n) and the laboratory's tendency estimated through the relative or straightforward difference between the laboratory's results and the target value, are also employed, depending on the proficiency test's purpose. Some authors aimed to validate the method, mainly related to repeatability and reproducibility [12, 22, 34].

Graphical methods

Requirement 10.1 of ISO 13528 defines the application of graphical methods. Considering this requirement, the PT provider can prepare graphs such as histograms of results or performance scores and Kernel density plots. Similarly, with the PT report, the graphs can show the results or participant's performance by codes, enabling each participant to know their own result. Although the graphs are important to evaluate and compare the laboratory results, the PT provider can determine if there is a need to review the criteria used to evaluate performance.

By using histograms, it is possible to evaluate the stability or degradation products of PT items [12], the percentage of

acceptable and unacceptable scores for each measurement in the submitted results from the laboratories participating in a PT study [13], and to evaluate the laboratory performance during a long-term PT study [17, 22]. Graphical methods for combining performance scores over several rounds of a PT scheme were found to be useful to allow detection of trends, such as the ratio between the uncertainty of the reference value and the assigned standard deviation ($u(x_{pt})/\sigma_{pt}$), employed in the calculations of the z-scores in a 10-year PT of multiple pesticide residues in olive oil. The evaluation of the trueness was also performed via recovery (%) versus concentration levels [22].

Another alternative graphical method widely used is S-shape charts to present the results obtained by laboratory participants with their expanded uncertainties for $k=2$. The results for each laboratory in the same plot are convenient to observe the performance of the participants [13, 19, 24].

The use of z-score graphs for an individual sample and performance evaluation (by the percentage of satisfactory results) was important for the PT provider to choose effective protocols for sample preparation, such as for the irradiation of samples included in a PT involving neutron activation analysis by the relative INAA and k_0-INAA methods for 16 elements in soil and plant samples [15], to correlate different methods used by the PT participants [22], to analyze the production of new products [16], and to rank the performance of the participants [24, 26, 27, 34].

To assess global performance in critical evaluation of a long-term PT of pesticide residues in olive oil, Generali et al. [28] used the average of the squared z-scores (AZ^2) to ascertain whether a laboratory had achieved the goal of detecting correctly at least 80% of the analytes of interest. This strategy is useful in the evaluation of methods to detect multiple pesticide residues in food.

The Youden plot is a very informative graphical method of studying the results when two similar PT items are tested in a round of a PT scheme. It can be useful for demonstrating the correlation (or independence) of results of different PT items and for guiding investigations into reasons for action signals (ISO 13528). Cordeiro et al. [19] applied Youden plots to evaluate the normal distribution of fipronil in egg samples. It was possible to visualize the samples with a bivariate confidence limit at 95% in a closed group, as an ellipse pattern in the center of the figure.

Thompson et al. [20], in a comparison between reproducibility standard deviations (SDR) derived from PT and collaborative trials, used log-transform the basic statistics (SDR and mass fractions) for regression and display purposes. The transformation provided a visual approach to a constant variance of the SDR x mass fractions. Simple regression was used to characterize each dataset over a wide range of mass fractions.

Box and whisker plots were used to indicate the variation among the interquartile results in a PT scheme. The whiskers indicated the maximum and minimum reported result [29]. In an olfactometric PT, Stöckel and co-authors [18] drew box plots of the recoveries per component to assess measurements' behavior. They also used recovery rates for each component per dosage, participant, and PT. The recovery results were converted into z-scores after logarithmic transformation to harmonize the participants' results from different components and PTs. Log-transformation before calculating z-scores is useful to establish near-symmetric distributions that are sufficiently close to normal to justify interpretation based on the normal distribution.

Another opportunity to visualize quantitative results for all samples is to use a chemometric approach, like principal component analysis (PCA). It was applied to perform multivariate analysis using quantitative data on phospholipid species to distinguish between five krill oil samples measured at different sites [12]. Becker et al. [33] used the Q score to evaluate participants' performance over consecutive rounds of 15 years of PT of total petrol hydrocarbon determination in soil. It is calculated as $Q=(x_{lab}-X_{ref})/X_{ref}$, expressed in %.

Conclusion

This review shows that the growth of concern over the reliability of analytical results is increasing the interest in PT and has thus increased the number of publications about this subject in the past few years. Although most of the papers found were about quantitative PTs, it is important to highlight the importance to discuss about qualitative PTs too, considering the different approaches to performance evaluation in this area.

Since different standards describe the statistical tools applied in each PT stage, considering the characteristics of each PT is necessary to choose the best strategies, for example, during the evaluation of an assigned value (characterization by a primary method in the laboratory, two different techniques or a consensual value between results reached from a PT after adequate statistical evaluation); during the evaluation of the results from the participants by different approaches (z-score, z'-score, Zeta-score); and the application of different graphical methods; among others.

Many PT studies discussed in this review were not found in the EPTIS database, so it was not possible to compare all PT protocols. However, these papers can support PT developments, even if some stages in some papers are not in precise conformity with ISO standards.

At present, due to the pandemic the accreditation bodies have adopted the use of remote audits to evaluate compliance with the ISO/IEC 17025 standard. Therefore, the technical requirements of that standard, which involve in-person monitoring of laboratory tests and other procedures, have been hampered. This means that the results of PT stand out for the ability to establish the technical competence of laboratories in support of remote audits.

Acknowledgements We acknowledge financial support from the São Paulo State Research Foundation (FAPESP, 2020/01238-4, 2018/26145-9) and the National Council for Scientific and Technological Research (CNPq, grant 308178/2018-1).

Funding Not applicable.

Availability of data and material Not applicable.

Code availability Not applicable.

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Valcarável M, Rios A (1994) Analytical chemistry and quality. Trac Trends Anal Chem. https://doi.org/10.1016/0165-9936(94)85055-0
2. Olivares IRB, Lopes FA (2012) Essential steps to providing reliable results using the analytical quality assurance cycle. Trac Trends Anal Chem. https://doi.org/10.1016/j.trac.2012.01.004
3. Taverniers I, De Loose M, Van Bockstaele E (2004) Trends in quality in the analytical laboratory. II. Analytical method validation and quality assurance. Trac Trends Anal Chem. https://doi.org/10.1016/j.trac.2004.04.001
4. ISO/IEC 17025:2017 General requirements for the competence of testing and calibration laboratories. Geneva
5. OECD (1998) Series on principles of good laboratory practice and compliance monitoring—number 1: principles on good laboratory practice. Paris
6. ISO 15189:2015 Medical laboratories—requirements for quality and competence. Geneva
7. ILAC (2020) Facts and figures. https://ilac.org/about-ilac/facts-and-figures/. Accessed Mach 2020
8. Olivares IRB (2016) Laboratory quality management. Editora Átomo, Campinas
9. ILAC (2014) Policy for participation in proficiency testing activities. https://ilac.org/latest_ilac_news/ilac-p9062014-published/. Accessed Mach 2020
10. ISO 13528:2015 Statistical methods for use in proficiency testing by interlaboratory comparison. Geneva
11. ISO/IEC 17043:2010 Conformity assessment-general requirements for proficiency testing. Geneva
12. Zailer E, Monakhova YB, Diehl BWK (2018) 31P NMR method for phospholipid analysis in krill oil: proficiency testing-a step toward becoming an official method. J Am Oil Chem Soc. https://doi.org/10.1002/aocs.12153
13. Yeltepe E, Şahin NK, Aslan N, Hult M, Özçayan G, Wershofen H, Yücel Ü (2018) A review of the TAEA proficiency test on natural and anthropogenic radionuclides activities in black tea. Appl Radiat Isotopes. https://doi.org/10.1016/j.apradiso.2017.10.011
14. Yarita T, Otake T, Aoyagi Y, Takasaka N, Suzuki T, Watanabe T (2018) Comparison of assigned values from participants' results, spiked concentrations of test samples, and isotope dilution mass spectrometric results in proficiency testing for pesticide residue analysis. J AOAC Int. https://doi.org/10.5740/jaoacint.17-0218
15. Esen AN, Haciyakupoglu S, Erenturk S (2017) Comparison of relative INAA and k0-INAA using proficiency test materials at ITU TRIGA Mark II research reactor. J Radioanal Nucl Chem. https://doi.org/10.1007/s10967-017-5669-0
16. Wiedner H, Riedl J, Maringer FJ, Baumgartner A, Stietka M, Kabrt F (2018) Production and characterization of a traceable NORM material and its use in proficiency testing of gamma-ray spectrometry laboratories. Appl Radiat Isotopes. https://doi.org/10.1016/j.apradiso.2017.09.025
17. Nakashima N, Duran EB (2018) Proficiency test exercises for particulate systems at CTBT radionuclide laboratories. Appl Radiat Isotopes. https://doi.org/10.1016/j.apradiso.2017.07.034
18. Stöckel S, Cordes J, Stoffles B, Wildanger D (2018) Scents in the stack: olfactometric proficiency testing with an emission simulation apparatus. Environ Sci Pollut Res. https://doi.org/10.1007/s11356-018-2515-z
19. Cordeiro F, Bratinova S, Karasek L, Buttinger G, Stroka J, Emteborg H, Seghers J, Robouch P, Emons H (2019) Can official control laboratories quantify reliably fipronil in eggs? Evidence from a proficiency testing round. Food Addit Contam. https://doi.org/10.1080/19440049.2019.1602885
20. Thompson M, Sykes M, Wood R (2019) Comparisons between reproducibility standard deviations (SDR) derived from proficiency tests and from collaborative trials: mycotoxins in food. Accred Qual Assur. https://doi.org/10.1007/s00769-019-01413-8
21. Dajay LC, Portugal TR, Climaco JC, Parcon MRV, Udarbe MA, Placio REE, Adona PE (2018) Establishment of proficiency testing programs in the Philippines. Accred Qual Assur. https://doi.org/10.1007/s00769-018-1363-3
22. Stefanelli P, Generali T, Girolimetti S, Barbini D (2018) Evaluation of the reproducibility standard deviation in the pesticide multi-residue methods on olive oil from past proficiency tests. Accred Qual Assur. https://doi.org/10.1007/s00769-018-1330-z
23. Thiex N, Carlson M, Kieffer R, Kieffer A, Eisenberg D, Bararshkov N, Ramsey C (2019) Evaluation of the use of microtracers™ in a proficiency testing program. J AOAC Int. https://doi.org/10.5740/jaoacint.18-0354
24. Aryana N, Ryana N, Ramadhaningtyas DP, Styarini D, Arisriawan Y (2019) First Indonesian proficiency testing using reference value from isotope dilution mass spectrometry method for benzoic acid, methyl paraben, and n-butyl paraben in sweet soy sauce. Accred Qual Assur. https://doi.org/10.1007/s00769-019-01398-4
25. Ferrini AM, Appicciafuoco B, Massaro MR, Galati F, Patriarca M (2019) Proficiency testing as an instrument to assess the analytical performance and the methods routinely implemented: the Italian experience for the screening of antibiotic residues in milk in the official control. Accred Qual Assur. https://doi.org/10.1007/s00769-018-1352-6
26. Kim H, Hwang E, Park J, Heo SW, Yim YH, Lim Y, Lim MC, Lee JW, Lee KS (2019) Proficiency testing for total mercury in oyster with a metrologically traceable reference value from isotope

dilution mass spectrometry: implications on laboratory practices using mercury analyzers. Accred Qual Assur. https://doi.org/10.1007/s00769-019-01379-7
27. Ebarvia BS, Dacuya A, Cabanilla SR, Mamplata NR (2019) Provision of proficiency testing for histamine mass fraction in canned tuna to improve the capability of chemical laboratories in the Philippines. Accred Qual Assur. https://doi.org/10.1007/s00769-018-1347-3
28. Generali T, Stefanelli P, Girolimetti S, Barbini DA (2019) Results of the 16th proficiency test on the determination of pesticide residues in olive oil. Accred Qual Assur. https://doi.org/10.1007/s00769-018-1329-5
29. Middlebrook KA (2019) A proficiency testing scheme to evaluate the effectiveness of laboratory sample reduction of a soil sample. Accred Qual Assur. https://doi.org/10.1007/s00769-018-1357-1
30. Ziegler E, Tirard A, Boubetra A, Bort M (2019) Acquisition of stability data for pesticides in water sample through proficiency tests. Accred Qual Assur. https://doi.org/10.1007/s00769-018-1339-3
31. Coleman MD, Smith TOM, Robinson RA, Stoffels B, Wildanger D (2019) Combining UK and German emissions monitoring proficiency testing data based on stack simulator facilities to determine whether increasingly stringent EU emission limits are enforceable. Accred Qual Assur. https://doi.org/10.1007/s00769-018-1354-4
32. Krismastuti FSH, Hamim N (2019) Designing a formulation of synthetic wastewater as proficiency testing sample: a feasibility study on a laboratory scale. Accred Qual Assur. https://doi.org/10.1007/s00769-019-01399-3
33. Becker R, Sauer A, Bremeser W (2019) Fifteen years of proficiency testing of total petrol hydrocarbon determination in soil: a story of success. Accred Qual Assur. https://doi.org/10.1007/s00769-019-01383-x
34. Leyva MJB, Batidas PJ, Muñiz VR, Ceballos MSG, Ponce VG, Aguilera MD, Grajeda CP, Navidad MMS, Flores MME, Ramírez COJA, Aguilar ZG, Huerta BG (2019) Measurement of organochlorine pesticides in drinking water: laboratory technical proficiency testing in Mexico. Accred Qual Assur. https://doi.org/10.1007/s00769-019-01403-w
35. Olivares IRB, Souza GB, Nogueira ARA, Toledo GTK, Marcki DC (2018) Trends in devolpments of certified reference materials for chemical analysis—focus on food, water, soil and sedment matrices, trends in analytical chemistry. Trac Trends Anal Chem. https://doi.org/10.1016/j.trac.2017.12.013
36. EPTIS (2020) About the database. https://eptis.org/about.htm. Accessed May 2020
37. Baker M (2016) Is there a reproducibility crisis? Nature 533:452–454
38. Lisinger TPJ, Pauwels J, Van Der Veen AMH, Schimmel H, Lamberty A (2001) Homogeneity and stability of reference materials. Accred Qual Assur. https://doi.org/10.1007/s007690000261
39. ISO Guide 35:2017 Reference materials -- general and statistical principles for certification, 4th edition
40. Thompson M, Ellison SLR, Wood R (2006) The international harmonized protocol for the proficiency testing of analytical chemistry laboratories (IUPAC technical report). Pure Appl Chem. https://doi.org/10.1351/pac200678010145
41. RSC (2016) z-scores, and other scores in chemical proficiency testing—their meanings, and some common misconceptions. AMCTB. https://doi.org/10.1039/c6ay90078j
42. Fearn T, Thompson MA (2001) A new test for 'sufficient homogeneity.' Analyst. https://doi.org/10.1039/b103812p
43. Dixon WJ (1960). Simplified estimation from censored normal samples. Ann Math Stat. https://www.jstor.org/stable/2237953
44. Rivera C, Rodríguez R (2020) Horwitz equation as quality benchmark in ISO/IEC17025.TestingLaboratory. Available at: https://pdfs.semanticscholar.org/d6d6/a38d1a9e01e526ca4e2b5b8d804670e5414f.pdf?_ga=2.149818485. 2108515611.1564572314–1074601812.1564572314. Accessed 30 July 2020
45. Horwitz W (1982) Evaluation of analytical methods used for regulations of food and drugs. Anal Chem. https://doi.org/10.1021/ac00238a765