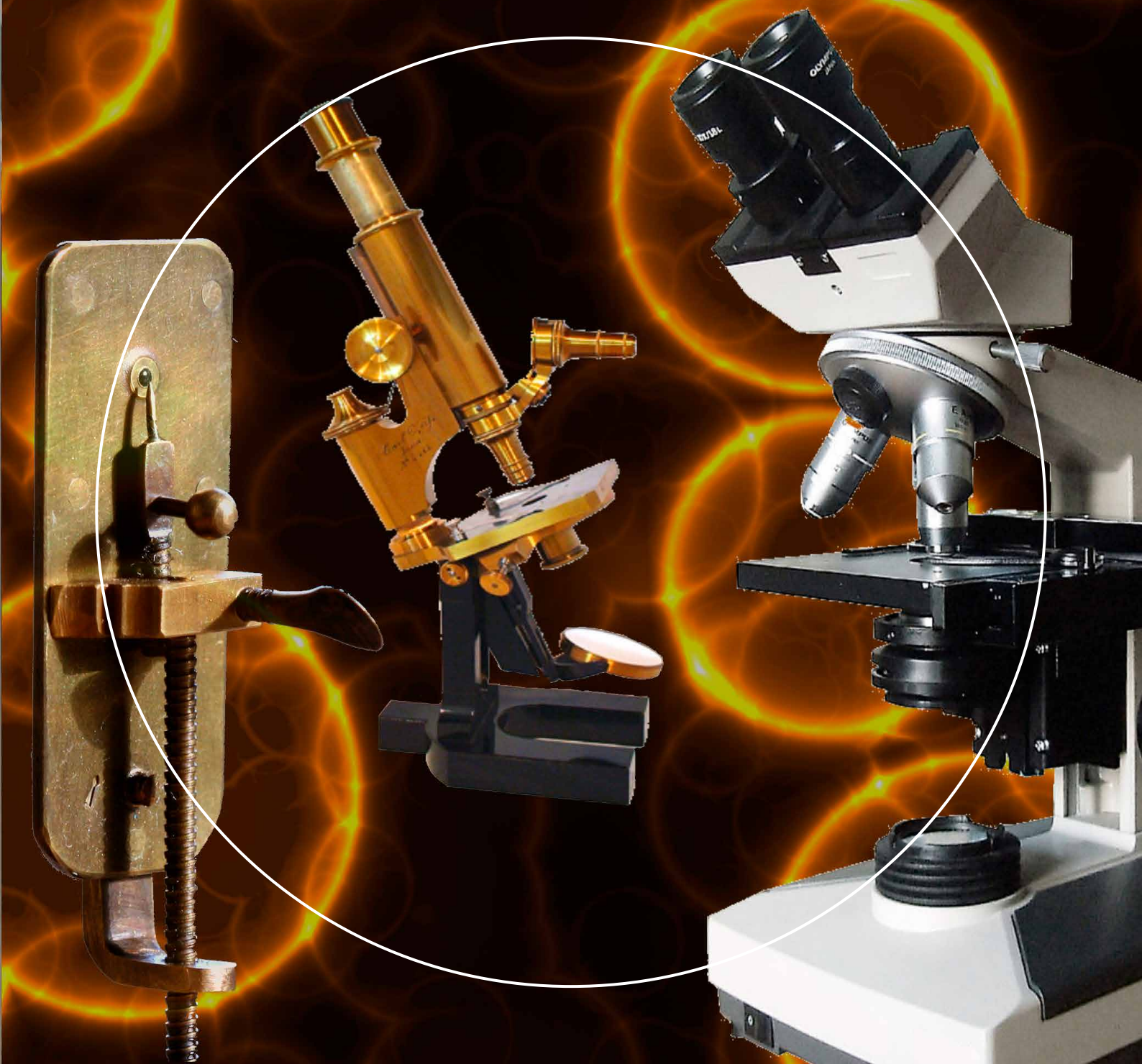
Quality assurance and control of methods to examine visually recognisable substances in feed and food

In cooperation with International Association for Feedingstuff Analysis, Section Feedingstuff Microscopy

L.W.D. van Raamsdonk, G. Frick, I. Ujčič Vrhovnik, M. Zadravec, J. Zegers, R. Krull-Wöhrmann, R. Weiss, G. van der Borg

WAGENINGEN
UNIVERSITY & RESEARCH



Quality assurance and control of methods to examine visually recognisable substances in feed and food

In cooperation with International Association for Feedingstuff Analysis, Section Feedingstuff Microscopy

L.W.D. van Raamsdonk, G. Frick, I. Ujčič Vrhovnik, M. Zadravec, J. Zegers, R. Krull-Wöhrmann, R. Weiss, G. van der Borg

This research has been carried out by Wageningen Food Safety Research, institute within the legal entity Wageningen Research Foundation subsidised by the Dutch Ministry of Agriculture, Nature and Food Quality (project number 122.72.595.01).

Wageningen, November 2022

WFSR Report 2022.006

WAGENINGEN
UNIVERSITY & RESEARCH

L.W.D. van Raamsdonk, G. Frick, I. Ujčič Vrhovnik, M. Zadravec, J. Zegers, R. Krull-Wöhrmann, R. Weiss, G. van der Borg, 2022. *Quality assurance and control of methods to examine visually recognisable substances in feed and food;* Wageningen, Wageningen Food Safety Research, WFSR Report 2022.006. 94 pp.; 3 fig.; 14 tab.; 169 ref.

Project number: 122.72.595.01
BAS-code: WOT-02-004-056
Project title: Microscopische expertise
Project leader: T.W. Prins

This report can be downloaded for free at https://doi.org/10.18174/580536 or at www.wur.eu/food-safety-research (under WFSR publications).

P.O. Box 230, 6700 AE Wageningen, The Netherlands, T +31 (0)317 48 02 56, E info.wfsr@wur.nl, www.wur.eu/food-safety-research. WFSR is part of Wageningen University & Research.

WFSR report 2022.006

Distribution list:

- International Association for Feeding stuff Analysis (IAG), Section Feeding stuff Microscopy (G. Frick, J. Vancutsem, R. Weiss, M. Zadravec, J. Dietz)
- Authors of the report (R. Krull-Wöhrmann, I. Ujčič Vrhovnik, J. Zegers, G. van der Borg)
- European Commission (EC; E. Thevenard, F. Schneegans, K. de Smet, F. Verstraete)
- European Union Reference Laboratory, Animal Proteins (CRA-W; O. Fumière, G. Berben, P. Veys, V. Baeten)
- European Union Reference Laboratory, Plant and Mycotoxins (WFSR; M. den Nijs, J.G.J. Mol, P. Mulder, T. de Rijk)
- Joint Research Centre, Geel (IRMM-JRC; C. von Holst, A. Boix-Sanfeliu)
- CEN Technical Commission 327 (WFSR; A. Verschoor)
- European Fat Processing and Renderers Association (EFPRA; M. Alm, S. Woodgate)
- European Feed Manufacturers' Federation (FEFAC)
- European Former Foodstuff Processors Association (EFFPA; A. van den Brink)
- International Fishmeal and Fish oil Organisation (IFFO; A. Jackson, I. Pike)
- Ministry of Agriculture, Nature Management and Food Quality (LNV; L.A.M. Claassen, A.D. Wentzel)
- Food and Consumer Product Safety Authority (NVWA; K. Zwaagstra, M. van Brakel, S. Danneel)

Contents

Expert panel **7**

Scope of the Guidance **9**

PART 1 Theoretical background **11**

1 Introduction **13**

2 Domain of visual examination methods for feed and food **15**

3 Quality guidelines and standards **17**

4 Quality parameters **19**

4.1 Quantitative methods 19
4.1.1 Recovery 20
4.1.2 Decision limit CCα: quantification limit 21
4.1.3 Selectivity/specificity 21
4.1.4 Uncertainty: count dispersal and weight uncertainty 21
4.1.5 Repeatability/reproducibility 23
4.1.6 Robustness/stability 24
4.2 Qualitative methods 24
4.2.1 Correctness (accuracy) 26
4.2.2 Detection capability CCβ: sensitivity 26
4.2.3 Selectivity/specificity 27
4.2.4 Repeatability and reproducibility: accordance and concordance 27
4.2.5 Robustness/stability 28
4.3 Estimation methods 28
4.3.1 IAG estimation model 28
4.3.2 Traces 29

5 Method development and QA/QC **31**

5.1 Method design 31
5.1.1 Material treatment and work flow 31
5.1.2 Examination 33
5.2 Identification 35
5.2.1 Identification support 35
5.2.2 Statistical tests 36
5.2.3 Probability of Identification 36
5.2.4 Strategies for validation of identification procedures 36
5.3 Conformity and equivalency of methods 37
5.4 Quality control of sample analyses 37
5.4.1 Workflow 37
5.4.2 Evaluation and documentation of the results 41

Literature **42**

Appendix 1 Confidence intervals **50**

Appendix 2 Definitions of terms **54**

Appendix 3 Expert systems performance criteria **58**

Appendix 4 Overview of scope of Guidelines **60**

PART 2 Validation 65

1 Background 67

1.1 Scope of the Guidance 67

2 Domain of visual methods for feed and food 68

3 Quality parameters 69

3.1 Quantitative methods 69
3.1.1 Recovery/bias 70
3.1.2 Detection limit CCβ: quantification limit 71
3.1.3 Selectivity/specificity 71
3.1.4 Repeatability and reproducibility 72
3.1.5 Uncertainty: count dispersal and weight uncertainty 75
3.1.6 Robustness/stability 76
3.2 Qualitative methods 77
3.2.1 Correctness (accuracy) 77
3.2.2 Detection limit: sensitivity 78
3.2.3 Selectivity 78
3.2.4 Repeatability and reproducibility: accordance and concordance 79
3.2.5 Robustness/stability 80
3.3 Identification 81
3.3.1 Validation of binary identification methods 81
3.3.2 General approach for establishment of identity and authenticity 83

4 Skewed data distribution 84

Appendix 1 Example cases 85

Expert panel

L.W.D. van Raamsdonk, Wageningen Food Safety Research, Wageningen, the Netherlands; board member IAG Section Feedingstuff Microscopy

G. Frick, Agroscope, Posieux, Switzerland; president IAG Section Feedingstuff Microscopy

I. Ujčič Vrhovnik, UL-VF, Institute of Food Safety, Feed and Environment, Ljubljana, Slovenia

M. Zadravec, Croatian Veterinary Institute, Zagreb, Croatia

J. Zegers, Nutreco Masterlab, Boxmeer, the Netherlands

R. Krull-Wöhrmann, Chemisches und Veterinäruntersuchungsamt Rhein-Ruhr-Wupper (CVUA-RRW), Krefeld, Germany; board member IAG Section Feedingstuff Microscopy

R. Weiss, Austrian Agency for Health and Food Safety, Vienna, Austria; board member IAG Section Feedingstuff Microscopy

G. van der Borg, Wageningen Food Safety Research, Wageningen, the Netherlands

Statistic consultancy for specific topics: H. van der Voet, Biometris, WUR, Wageningen, the Netherlands; C. von Holst, JRC, Geel, Belgium

Scope of the Guidance

This Guidance presents an elementary overview of the validation and application of methods to examine visually recognisable substances in feed and food. Although quality parameters and criteria for visual inspection methods have been taken frequently from other major disciplines such as analytical chemistry and microbiology, it is necessary to acknowledge that visual examination is a completely different discipline, which need different approaches. Basic principle is the examination of undesired substances or ingredients occurring as visible units of extremely larger size and dimensions than chemical molecules. As consequence, the statistical background and physical distribution are principally different from those molecules affecting the design and validation of visual examination methods.

This Guidance will present and discuss specific application of relevant quality parameters to visual detection methods including microscopy and a framework of dedicated sets of quality parameters for the domain of visual monitoring methods will be given, separate for quantitative, qualitative and estimation methods. Elements for the design of visual examination methods are presented and discussed in relation to quality parameters.

Part 1 will provide the theoretical background and principles of quality assurance and control for visual inspection. A separate Part 2 consists of an overview of the relevant parameters and procedures for application, which can be used as template for a Standard Operational Procedure for validation in a quality management system.

PART 1 Theoretical background

1 Introduction

Safety is an intrinsic part of the production of feed and food. A long history exists of measures and control procedures. In order to cover the broad range of hazards and in particular the regulated and restricted components, monitoring methods are developed and in use from several disciplines. A set of four basic safety domains can be distinguished. These include analytical methods in the area of **B**iology, e.g. GMO detection, prions, processed animal proteins and toxic seeds, in the area of **C**hemical compounds, e.g. pesticides, antibiotics, growth promotors, heavy metals and process contaminants (e.g. dioxins), and in the area of **M**icrobiology such as bacteria, zoonoses, pathogens and viruses. Analytical chemistry is the most prominent discipline in terms of the range of methods and quality assurance standards. At the other end of this spectrum hazards exist in the area of **P**hysics, e.g. radionuclides, (micro-)plastic, inorganic components (Besatz) and packaging material. These elements of what can be indicated as the **BCMP** cocktail are laid down in several EU directives and regulations (Directive 2002/32/EC, Regulation (EC)1831/2003, Regulation (EU) 999/2001 a.o.) and reviewed in an FAO/WHO report (FAO and WHO, 2019). Some of these domains are well addressed, whereas others are poorly understood. The latter holds in particular for the domain of physical hazards.

An increasingly important area is authenticity and identity analysis of materials and ingredients in food and feed. Results of these examinations are the basis for label control, track and trace and transparency in food production. These mechanisms support a targeted monitoring of feed and food safety. In a range of cases elements of biological and processing background are part of the description of ingredients, e.g. in the Feed Catalogue (Regulation (EU) 68/2013, European Commission 2013e, amended according to Regulation (EU) 2017/1017, European Commission 2017b).

Monitoring of biological and physical hazards is partly based on visual observations, including but not limited to microscopy. In terms of actual monitoring, visual methods include targets or contaminants such as bone fragments or other particles of animal origin, plant seeds, spore bodies of moulds, sclerotia, packaging material, and "Besatz". Identification of legal ingredients and composition analysis is part of the domain of visual examination as well.

It has to be emphasized that methods for visual or microscopic examination are based on other assumptions and principles for method design and quality assurance than those used in analytical chemical methods. This has four reasons:

1. Inhomogeneity plays a different role because the units to be detected are large considering their logic visibility. The resulting minimum size of the detection unit is in principle larger than 1 µm, which is many times larger than chemical molecules. The a-priori presumed consequence is a larger variation among replicate samples compared to chemical methods at comparable contamination levels, which would eventually lead to high levels of measurement uncertainty. This situation prevents to have duplicate samples as intended in most chemical analyses.
2. The size of the analysis sample is typically in the range of grams, from 10 grams for analysis of animal proteins (microscopic) up to 500 grams for undesirable substances (macroscopic). The smaller the test portion size, the larger the probability of false negative results and the larger the uncertainty of the actual level of contamination in the test portion.
3. The sensitivity of visual methods is in the ppm range or higher (25 ppm = 0.0025%). Levels in the ppb range can be assumed to be below the detection level.
4. There is no data processing system that detects the targeted substance(s) and provides their identity, such as a mass spectrometer, PCR equipment or Biacore reader, in the vast majority of cases for visual examination. The detection and identification are performed by the microscopic technician based on knowledge and experience. In all cases the performance of the method (the protocol) and the performance of the technician (the expertise) has to be addressed separately. Both aspects contribute to the quality of a visual method, but in their own way.

Much investments has been given to the development of procedures for quality assurance and control, with emphasis on analytical chemistry. An overview of standards as background documentation is given in Appendix 4. The situation that visual detection is principally based on units which are much larger than chemical molecules has major implication on the applicability of parameters for method validation and performance. Nevertheless, in a range of cases applications of quality parameters for visual methods were still adopted from chemical standards, which can result in confusing situations. An introduction to the cause and background to develop this Guidance has been published (van Raamsdonk et al., 2022a). A general framework for quality assurance for visual examination methods should be explored and defined, presenting quality parameters in the domain of visual research.

2 Domain of visual examination methods for feed and food

The set of methods of the domain of visual inspection covers a range of magnifications of the target under study. The scope of visual methods includes anything that is or can be made visible. The domain includes three levels of particle size, each demanding its own approach and equipment (Table 1).

Table 1 *Overview of possible methods, arranged according to purpose: detection of undesirable substances at three levels of particle size, estimation and identification.*

Equipment and magnification	Type of matrix	Particle size	Example targets
Quantitative			
Visual, non-microscopic: none or magnifying glass; 1 to 8x	Unground or raw materials	Preferably larger than 1 mm	• Unground feed materials, whole kernel cereals, bird feed; contamination with seeds or sclerotia, 'Besatz' • Soil, fertiliser, manure; microplastic over 1 mm, macroplastic • Former food products, compound feed; packaging materials • Hay; toxic plants • Living or dead insects, their damage
Semi-quantitative or qualitative			
Macroscopic: binocular or stereo microscope; 8 to 64x	Ground materials; coarse	200 – 1000 µm	• Compound feed; animal by-products, coarse fraction • Feed ingredients; identification, coarse material • Blood plasma (visual marker for haemoglobin activity) • Living or dead insects, mites, their eggs, damage and dejections • Mould infestations
Qualitative			
Microscopic: compound microscope; 100 – 400x	Powder or meal, fluids with cells or particles; fine	10 – 200 µm	• Compound feed; animal by-products, fine fraction • Compound feed; ground botanic impurities • Compound feed, food; microplastic • Oil: animal proteins or other inclusions • Herbs or spices; adulteration • Milk; somatic cell count • Spores from moulds, yeast
Estimation methods			
	All types	All sizes	• Feed; estimation of botanic composition • Undesirable substances; traces of ground *Brassica* seeds • Cereal flour; traces of biomarkers for allergens
Identification of matrix material			
	All types	All sizes	• Compound feed; identification of minerals (spot tests) • Honey or other matrix; identity pollen • Flour or meal; starch identification • Milk powder • Rice; identity • Pine nuts; identity

Physically selected and separated material can in principle be quantified (counted and/or weighed). Usually weight percentages will be taken as final measure in order to report documentation on possible excess of legal limits. Directive 2002/32/EC includes a range of limits on undesirable impurities of botanic origin (seeds of *Datura*, *Ambrosia*, *Brassica* (mustard) and some other species) and ergot sclerotia. For most of these impurities, IAG section Feedstuff microscopy has developed methods for monitoring (Table 2A). In addition, methods for visual inspection has been published by other organisations (Table 2B).

Table 2A *Methods developed by the International Association of Feedingstuff Analysis, Section Feedingstuff Microscopy.*

IAG-A1	Sample Preparation for the Macroscopic and Microscopic Analysis
IAG-A2	Method for the Identification and Estimation of Constituents in Animal Feedingstuff
IAG-A3	Determination of *Datura* spp. In Animal Feedingstuff
IAG-A4	Method for the Determination of Ergot (*Claviceps purpurea* Tul.) in Animal Feedingstuff
IAG-A5	Method for the Determination of Fruits and Seeds of *Ambrosia* spp. In Animal Feedingstuff 2019
IAG-A6	Method for Determination of Castor (*Ricinus communis* L.) seed husks in Animal Feedingstuff
IAG-A7	Method for the Determination of Stone Shells in Animal Feedingstuff
IAG-A8	Method for the Determination of poisonous plants in roughage
IAG-A9	Determination of Rice chaffs (*Oryza sativa*) in Feedingstuff

Table 2B *Other methods for visual examination.*

Regulation (EC) 152/2009, Annex VI, Part 2.1	Detection of animal proteins in feeds and feed ingredients
Regulation (EU) 1308/2013, Annex II, Part I	Measurement of rice in the framework of custom taxes
Regulation (EC) 401/2006	Methods for sampling and analysis of mycotoxins in foodstuffs (Annex II describes visual inspection)
ISO 11746:2012	Rice – Determination of biometric characteristics of kernels
ISO 3632-2:2010	Spices — Saffron (*Crocus sativus* L.) – Part 2: Test methods
ISO 5061:2002	Animal feeding stuffs - Determination of castor oil seed husks – Microscope method
EN-ISO 13366-1:2008	Milk – Enumeration of somatic cells – Part 1: Microscopic method (Reference method)
EN-ISO 658:2002	Oilseeds – Determination of content of impurities
CEN 15587:2018	Cereal and cereal products – Determination of Besatz in wheat (*Triticum aestivum* L.), durum wheat (*Triticum durum* Desf.), rye (*Secale cereale* L.), triticale (*Triticosecale* Wittmack spp) and feed barley (*Hordeum vulgare* L.)
EN 16378:2013	Cereals – Determination of impurities content in maize (*Zea mays*, L.) and sorghum (*Sorghum bicolor,* L.)
EURL myco-/plant toxins method 006	Determination of ergot sclerotia (*Claviceps purpurea* Tul.) in whole kernel cereals by visual screening
DIN 10760:2002	Determination of the relative pollen content of honey
VDLUFA MB VI-M 9.2	Zählung somatischer Zellen in Rohmilch: Mikroskopische Zählung somatischer Zellen (Referenzverfahren)
VDLUFA MB III 30.2	Bestimmung von Mutterkorn in Futtermitteln
VDLUFA MB III 30.3	Bestimmung von *Datura* ssp. In Futtermitteln
VDLUFA MB III 30.5	Bestimmung von Rizinus-Samenschalen
VDLUFA MB III 30.7	Identifizierung und Schätzung von Bestandteilen
VDLUFA MB III 30.8	Bestimmung von *Ambrosia Astemisiifolia* L.
VDLUFA MB III 30.9	Bestimmung von makroskopisch/mikroskopisch erfassbaren Fremdbestandteilen in Futtermitteln
van Raamsdonk et al., 2012	Examination of packaging materials in bakery products. A validated method for detection and quantification
Amato et al., 2017	Gravimetric quantitative determination of packaging residues in feed from former food
Veys et al., 2018	Isolation of insect material from feed using double sedimentation
Marchis et al., 2021	Gravimetric quantitative validation of botanic impurities in feed
van Raamsdonk et al., 2022	A validated method for detection and quantification of packaging material in candy syrups

Some seeds such as mustard seeds have to be monitored when ground as well. In those cases quantification is not possible and the presence of "traces" has to be reported. The possibility to quantify animal proteins in compound feeds has been explored, but for several reasons this appeared to be unreliable. Major issues appeared to be the translation from a two-dimensional view to a three-dimensional estimation of the size of the particles, the different specific densities of the materials, and the usual overestimation of materials with a low share in the composition (Veys et al., 2010). Generally, in all those cases where fragments cannot be physically selected, i.e. with particle sizes below approximately 1 mm or 1000 µm, methods can be assumed to be restricted to qualitative results, using microscopic methods.

3 Quality guidelines and standards

The major framework for quality assurance in food safety among other areas of application is the International Standard ISO/IEC 17025:2017 ("General requirements for the competence of testing and calibration laboratories"). Besides a range of measures on organisation and administration, chapters 7.2 (Selection, verification and validation of methods) and 7.6 (Evaluation of measurement uncertainty) are of particular importance for the current Guidance. The ISO Standard requires to apply one or a combination of several of the following procedures (ISO/IEC 17025:2017: Chapter 7.2):

a. Evaluation of trueness and precision based on reference material.
b. Assessment of factors influencing the results.
c. Testing the robustness of the method.
d. Comparing the results with those as achieved by other methods.
e. Interlaboratory comparisons.
f. Evaluation of measurement uncertainty.

ISO/IEC 17025:2017 Chapter 7.6 provides further information on the establishment and evaluation of measurement uncertainty. The current Guidance are based on the principles of the ISO Standard and discusses the applications and limitations of its measures for visual monitoring. Some issues concern the use of measurement uncertainty as general principle, the interpretation of duplicate analyses and the lack of quantitative results for microscopic methods. The principles of the organisation of Interlaboratory Comparisons (item e.) are laid down in ISO/IEC 17043:2010. The scope of this ISO Standard includes the evaluation of the performance of a method (collaborative trials). A set of actions is listed which cover the procedure of proficiency testing. With respect to visual methods, one of these actions is of particular importance: "to operate the data processing system" (ISO/IEC 17043:2010: paragraph 4.2.4, item g). A data processing system for visual observations need a diverging interpretation than understood in analytical chemistry. Additional requirements for training and education, with emphasis on proper recognition and classification, are required for technicians applying visual methods. This aspect would fit in the requirements as set out in paragraph on the necessary qualifications of personnel in both ISO Standards (ISO/IEC 17025:2017: paragraph 6.2; ISO/IEC 17043:2010: paragraph 4.2). ISO Standard 13528:2015 provides the framework for statistical analysis of interlaboratory comparisons. One section is devoted to qualitative tests (section 11).

Usually quality parameters are taken from the domain of analytical chemistry for application in the discipline of visual research. A lot of standards are focusing on elements of chemical analysis or are largely restricted to quantitative methods (examples are ISO, 1994; Horwitz, 1995; AOAC, 2002; ISO, 2005; SANCO, 2009; AOAC, 2016; Regulation (EU) 2021/808; see Appendix 4). The scopes of the standards are illustrated by the absence of terms applying to qualitative methods, by results indicated in terms of measurements, by the explanation of parameters for specific chemical methods, and by using examples taken from this domain. These parameters are intended to characterise "measurement results", suggesting that these criteria are initially intended for quantitative methods. In particular, parameters such as repeatability, reproducibility and linearity apply to quantitative and/or chemical methods. Several Standards in the domain of microbiological testing implicitly mention the difference among quantitative and qualitative methods (Feldsine et al., 2002; FDA, 2015; EPA, 2016). These standards and approaches can be used as a starting point for developing dedicated sets of quality parameters for visual methods distinguishing qualitative and quantitative methods.

Regulation (EU) 2021/808 contains a table with the minimal requirements separately for quantitative and qualitative method validation. This table is basically shown in Table 3, complemented with parameters taken from other sources. Definitions of parameters and other terms are listed alphabetically in Appendix 2.

Regulation (EC) 401/2006 and Regulation (EU) 2021/808, among others, discriminate between screening methods and confirmation methods for chemical compounds. Screening is intended to provide information on the presence of an analyte at a certain level without precise identification, which means that often a group of

analytes is detected by means of e.g. bio-analytical methods with two possible results: negative or suspect. That certain level, the detection capability or CCβ, is related to the performance of the intended confirmatory method. The false negative rate of the screening method should not exceed 5% ($\beta < 0.05$). The (usually quantitative) confirmation method needs to be capable of precise identification, for instance by high resolution mass spectroscopy. Besides specificity, decision limit (CCα), trueness/recovery and precision apply as performance parameters for confirmation methods. It is not current practice to differentiate between screening and confirmation methods for visual monitoring methods. This situation complicates the optimal choice of performance parameters for visual examination methods.

Table 3 *Overview of principal quality parameters, organised in two different groups for Accuracy and Precision. The table is largely based on Regulation (EU) 2021/808 with additions from supplemental sources. This overview does not include the translation to the domain of visual methods. S: screening, C: confirmation*

Parameter	Quantitative	Qualitative (ordinal)	Reference	notes for qualitative methods
Accuracy				total of sensitivity and specificity
Trueness/recovery	S, C		*A, B, C*	
Detection limit CCβ	S	S	*A, B*	sensitivity
Decision limit CCα	C	C	*A*	
Selectivity/specificity	C	C	*A, B, C*	Ref. *C*: exclusivity
Precision				
Repeatability/reproducibility	S, C		*A, B, C*	Ref. *D*: accordance and concordance
Robustness/stability	S, C	C	*A*	

A: Regulation (EU) 2021/808, repealing Commission Decision 2002/657/EC

B: AOAC, 2002

C: AOAC, 2016

D: Langton et al., 2002

Depending on the purpose (detection of a contaminant or establishment of a composition / identity) and the type of method (quantitative or qualitative), only a selection of the different quality indicators can be applied to visual research. The presence of a target in terms of single units has to be evaluated in a different way than a signal caused by trillions of molecules [1]. The presence of a visually detected particle means a region with a concentration or condensation of the target of 100% in terms of chemical molecules, whereas the material adjacent to that particle shows no contamination, which can be interpreted as a major inhomogeneity. The situation for qualitative tests has its consequences for the calculation of 95% confidence intervals in proficiency testing (Wehling et al., 2011; Macarthur and von Holst, 2012; Appendix 1), for parameters such as repeatability and reproducibility, to be replaced by concordance and accordance (Langton et al., 2002), and for finding a way to indicate first and second order errors (Regulation (EU) 2021/808, repealing Commission Decision 2002/657/EC; European Union, 2021). Situations comparable to the qualitative detection of particles (final result: presence or absence) have been evaluated for immunoassays (ruminant detection: van Raamsdonk et al., 2012, 2015). Such studies provide the first elements of a framework of quality parameters for visual inspection.

[1] "Units" in the sense as applied in this guideline are different from the units (S.I. units) as used in other guidelines (e.g. Eurachem QUAM:2012, page 9). See Appendix 2 for definitions.

4 Quality parameters

Visual methods, as applied to the types of materials with larger units (i.e. non-microscopic), result in *quantitative* results: primary results are counts in most cases with the derived results in weight percentages. Such results would allow to derive criteria from general Standards. The quantitative methods will therefore be discussed at first. A range of *semi-quantitative or qualitative* methods produce principal results in terms of a measurement or numerical value, translated to presence or absence of the target or contaminant (Table 1). Some methods result directly in a final conclusion on absence or presence. The few Guidelines considering qualitative methods do not distinguish the different ways in which a final conclusion can be reached, either semi-quantitative or fully qualitative, and qualitative methods will be discussed here as one group as well. In addition to this, a special type of visual examination has to be addressed, which is the estimation of the *composition or identity* of a sample. This is in particular applied to compound feeds and a dedicated procedure for estimation of the shares of the different ingredient is designed, with a model for establishing upper and lower confidence limits. Where applicable, dedicated terms will be introduced pertaining particularly to the domain of visual monitoring methods. These are included in Appendix 2, printed in **bold**.

Besides the monitoring of a target, contaminant or undesirable substance in terms of accuracy and precision, issues of identification will be addressed. Basically, units such as seeds, sclerotia and fragments of animal origin have to be identified in order to meet the demands of specificity (or exclusivity). Identification of subjects applies to a much broader domain, including the sample matrix and ingredients of composed materials.

4.1 Quantitative methods

The situation that visually recognisable contaminants are occurring in units (sclerotia or fragments thereof, seeds, Besatz) implies that the resulting principal measure is a number per portion of the matrix material. This will usually be translated to a weight percentage (w/w) by weighing the selected units, which is necessary in the view that a range of visually detected targets have weight limits (e.g. in Directive 2002/32/EC). Two types of results can be identified for this group of methods, one basic result in terms of counts, and one derived measure as weight percentage (w/w). Counts are principally zero or natural numbers above zero and will follow a Poisson distribution with an expected average λ and variance $\sqrt{\lambda}$ having an equal value. Weights can take any rational number and these results would follow normal distributions. Normal distributions are described by an average ($\mu = \sum(x_i)/n$) and a standard deviation ($\sigma = \sqrt{\frac{\sum(x_i-\mu)^2}{n-1}}$), with x_i = observation per sample, μ = average and n = number of observations. At lower contamination levels the distributions will be skewed, i.e. showing long right tails and will be best described following a lognormal distribution. Both types of results act differently, but are quantitively described. An average will be subjected to the performance parameters of the group Accuracy and a variance to the parameters of the group Precision.

A typical validation report could be based on the derived measure (weight) to indicate recovery, detection limit, specificity, uncertainty and robustness, with a relevant application for these parameters, and additional reference to the principal data: the counts. Only a small number of units is present at lower contamination levels. This results in a relatively large inhomogeneity, which influences the measurement uncertainty. This situation is completely different from chemical molecules, which are still present in large numbers at very low levels.

Inhomogeneity has been documented since 2016 for the detection of *Ambrosia* seeds, *Datura* seeds and Ergot sclerotia at WFSR. Samples of approximately 2 kg were divided in four equal portions and individually

analysed. Large inhomogeneity among the subsamples was found in terms of weight (mg/kg), especially at lower contamination levels (Appendix 1). In a situation where one subsample contains one seed or sclerotium (in general: one unit) and none is present in the other three subsamples, the homogeneity cannot be improved. Still, after calculation of a measure such as a relative standard deviation, the inhomogeneity appears to be large, especially when expressed as relative weight (w/w). Besides an effect for the application of specific criteria, subsamples cannot be considered true duplicate samples as considered for chemical analysis (Thompson and Wood, 1995; Reiter et al., 2011), for their different contamination level. In the absence of duplicate results, repeatability cannot be calculated, which means that reproducibility has to be calculated in a modified way.

It might be needed to establish or verify the identity of the selected material. If some units of the selected material are going to be subjected to an identification procedure, the entire amount of the selected material need to be counted and weighed beforehand. This would prevent that a possible destructive nature of a procedure could result in a lower recovery.

In several visual inspection methods resulting in a quantitative result, intermediate counts are not made. This especially applies to packaging material in former food material. These methods are still based on visible units, and it can be assumed that the inhomogeneity among subsamples would follow a relationship with levels of contamination comparable to that of other visual methods such as seeds and ergot sclerotia in whole kernel samples, instead of inhomogeneity levels as found after chemical analysis.

4.1.1 Recovery

Usually legal limits for undesirable substances are set at levels between 50 and 1000 ppm (mg/kg). According to AOAC (2002) and Codex Alimentarius (2004), upper and lower confidence limits should be set at 85-110% (100 ppm) and 90-108% (1000 ppm) respectively for accuracy / recovery at those levels. The deviation below or above the ideal value of 100% for recovery must be interpreted differently than with chemical research results. A recovery higher than 100% means that material has been selected that resembles the undesirable substance (seeds, ergot sclerotia, packaging material), but has a different source. This problem is referred to as selectivity/specificity and is addressed as another parameter. A recovery of less than 100% means that not all of the material present has been detected, recognized and selected, and / or that the selected units are damaged and very small fragments are left in the sample material. These are aspects of sensitivity. With a sufficient expertise of the operator to recognise mimics of the visual target, usually a recovery percentage of 100% or lower can be expected. Only small deviations in the weighing (measurement error in the analytical balance, increasing moisture in the target particles after selection from the matrix due to air humidity) can result in a small excess. These aspects will be discussed further in the paragraph on stability.

Correction for a moisture content of 12% is not necessary for the botanic impurities as listed in Directive 2002/32/EC, although required in that Directive. Seeds and ergot sclerotia will absorb moisture from or release to the surrounding matrix material (grains or other seeds) up to a final equalised level. An identical correction for the moisture level of the numerator and of the denominator in a quotient is not necessary. The same principle of absorption (water and / or fat) applies to paper and carton as parts of packaging material. In these and other cases not regulated by Directive 2002/32/EC a correction for 12% moisture is not required. Non-organic materials such as plastic, metal foil, glass and certain types of Besatz can be assumed to be incapable of water absorption.

The applicable limits depend on the circumstances of the validation study. Recovery of well-defined entities such as seeds and ergot sclerotia from samples that are individually spiked and fully examined should be close to the spiked number or within limits of statistical significance (95-105%). Subsamples might have a larger variation, as pointed out in Appendix 1. Proposed limits would be 34-166% (200 ppm) or 66-134% (1000 ppm) depending on the contamination level and type of matrix. Further documentation is given in the paragraph Uncertainty. Packaging material shows a wider range of backgrounds and possibly a higher diversity for recovery. A wider confidence limit might be applied.

4.1.2 Decision limit CCα: quantification limit

The use of parameters adjusted for the validation of chemical analysis methods can result in values without practical application in visual research. An example is the quantification limit of 0.41 mg/kg for *Ambrosia* seeds established from examination of samples of 500 gram each, which would mean that even less than 10% of one *Ambrosia* seed can be detected (WFSR validation study, 2012), which is by principle impossible when only complete units can be present. Weight is a derived result from the number of selected units. A calculated detection limit in terms of weight can therefore only be a theoretical indication. Expressed in number of units, and in the assumption of a well-trained technician, a detection limit of one (1) unit per amount of the investigated portion can be assumed to be reasonable in terms of numbers (counts) for visual methods.

Proposed criterium for LOD: 1 unit, which can be recalculated to a limit in mg/kg. If more than 1 kg will be investigated, a lower limit can be argued. As example: detection of 1 unit in 2 kg can be used to conclude an LOD of 0.5 unit per kg.

4.1.3 Selectivity/specificity

Monitoring methods should be capable of distinguishing between the target material and any other confusing material, which can mimic the target material or can confuse a correct identification of particles. Validation studies and proficiency tests should include an experiment dedicated to establish this parameter based on samples with two contaminants: the target and a mimicking non-target. The resulting recovery can be compared to the recovery without the mimicking non-target, and should not exceed the limits as set for recovery. Examples are the addition of a mimicking seed as non-target for *Ambrosia* in bird feed, or brand (cereal grains infected with mould) as non-target for ergot sclerotia in whole kernel cereals, or non-plastic microparticles in a method for monitoring microplastic. Besides identifications issues, laboratory contamination can result in false positives. The regular work flow (see paragraph 5.1.1 in Part 1) should include the examination of negative control samples at certain time intervals, at least biannually. The size of the particles which can be manually selected and handled would allow to prevent large scale carry-over among test samples.

Proposed limits: specificity is one of the causes of a deviating recovery, and the limits for recovery remain.

4.1.4 Uncertainty: count dispersal and weight uncertainty

Measurement uncertainty is a principal element in quality assurance (International Standard ISO/IEC 17025:2017, Chapter 7.6). A Range of guidelines and standards have been published providing definitions and explanations for this parameter (Pocklington, 1990 (IUPAC); Thompson et al., 2002 (IUPAC); AOAC, 2002b; Codex Alimentarius, 2004; Ellison and Williams, 2012 (Eurachem QUAM:2012); Bettencourt da Silva and Williams, 2015 (Eurachem STMU:2015); among others). The most comprehensive guidance is Eurachem QUAM:2012.

Several causes can be distinguished for the uncertainty of quantitative results, such as sampling, inhomogeneity after subsampling, the condition of the sample, the laboratory conditions, nature of the method, reagent purity, the quality of reference materials, equipment and instrumental effects, noise level in blanks, and operator influence (Codex Alimentarius, 2004; Eurachem QUAM:2012; Korol et al., 2015). Bias is not assumed to be part of the measurement uncertainty according to AOAC (2002b). As example of the nature of the sample, mycotoxins can occur in clusters or nuggets in a sample. Homogenization is important and specific procedures are designed for chemical analysis (Reiter et al., 2011).

The range of relevant factors for uncertainty in visual examination differs from that in chemical analysis. This is not only caused by the unit size for detection, but also by the low relevance of reagent purity and noise levels in blanks, excellent stability, and the situation that the expertise of the technician is a pivotal factor in the virtual absence of analytical equipment. In the particular case of plant seeds or sclerotia, these particles can well be preserved and described, based on a long history of botanical knowledge. Inhomogeneity remains as major factor in uncertainty for these samples (van Raamsdonk and van der Voet, 2022). If the usual procedure will be followed, the relative standard deviation over results of subsamples under

reproducibility circumstances (RSD_R) would be the principal indication of the measurement uncertainty. This value can be as high as 173% at low levels of contamination of seeds and Ergot sclerotia in whole kernel materials (Appendix 1). An important factor is the size of the units: with particle sizes below approx. 200 µm, which are typical for animal proteins as example of a microscopic examination method (Part 1, Chapter 4.2), sample material can sufficiently be homogenised, as demonstrated in numerous proficiency tests by applying the technique of step-wise dilution. The consequence is that undesirable substances other than seeds, well defined units with sizes in the mm range, would follow other uncertainty distributions. Packaging material consists of particles with a wide diversity of backgrounds (paper and board, plastics, aluminium foil, etc.) and sizes. The types of matrices of former food products, ranging from granular particles of different size (bakery by-products) to semi-fluids or fluids (candy syrup), are an additional factor in the extent to achieve homogenisation. Semi-fluidity of the matrix can be one of the factors to reach a reasonable homogenization (van Raamsdonk et al., 2022). It can be concluded that a combination of unit size and type of matrix will result in different levels of inhomogeneity after subsampling, ranging from whole kernel feeds and undesirable substances (seeds) with large inhomogeneity to (semi-)fluid matrices and/or small units with lower or reasonable levels of inhomogeneity.

In this Guidance the approach of an expression relative to the contamination level will be used (Ellison and Williams, 2012 (Eurachem QUAM:2012): section 8.2.6, Rule 2 (page 27), section 8.2.9 (page 28)). Analogue to the procedure of AOAC (2002b) and Horwitz and Albert (2006), the expected RSD_R and the upper limit of the expanded weight uncertainty U^* can be calculated from the datasets used in Appendix 1 for large undesirable substances and ergot sclerotia in granular matrices under the assumption of an average unit length of 5 mm. The relationships are expressed by the following equations (see also Appendix 1):

$$RSD_R = C^{-0.41} \quad \text{and} \quad U^* = 2 \cdot C^{-0.41}$$

This is the first derivation ever of an equation for the HORRAT value in the visual domain. Although with a still small basis (four datasets) and only applicable to unit lengths of approximately 5 mm, a first attempt to model the relationship between contamination level and uncertainty is worthwhile to be made. Table 4 shows the calculated RSD_R and the expanded uncertainty U^*.

Under the alternative provisions, the term "measurement uncertainty" is not recommended to be used for quantitative results of visual methods in order to avoid evaluation of levels of uncertainty along criteria commonly used in chemical analysis (AOAC, 2002b; Codex Alimentarius, 2004; Ellison and Williams, 2012 (Eurachem QUAM:2012); etc.). Instead, the term weight uncertainty is proposed. At this moment this term is documented for undesirable substances with an average unit length of 5 mm in dry granular sample material.

Table 4 *Maximum values for uncertainty of results expressed as weight% (w/w) for undesirable substances in dry granular material after subsampling, related to the level of contamination (mg/kg) under the provision of unit lengths of approximately 5 mm.*

Spike	concentration /mg/kg	Expected RSD_R	Expanded uncertainty (w/w)
0.0025%	25	77	not applicable
0.005%	50	58	not applicable
0.01%	100	44	87
0.02%	200	33	66
0.05%	500	23	45
0.1%	1000	17	34

A preferred alternative procedure for validation includes the selection, counting and weighing of undesirable substances from the sample, reintroduction of the selected material in the matrix, a waiting period of days up to a week, followed by a second selection, counting and weighing. In this procedure uncertainty is a minimal factor for unit counts, but might play a larger role for evaluating unit weights. Inhomogeneity as major factor for measurement uncertainty is basically absent when following the procedure of reintroduction

of the target material. This approach can be described as examination under reproducibility conditions. This will be discussed further in the next paragraph.

4.1.5 Repeatability/reproducibility

The calculation of repeatability and of reproducibility depends on the situation of the specific case. These circumstances include the spectrum of levels of inhomogeneity as discussed in the previous section (4.1.4). In order to achieve a validation plan suited for the specific situation, two extra parameters need to be considered. These are the amount of the sample material which would or would not allow to make subsamples, and the nature of the method, which can be destructive for the sample matrix preventing the option of reintroduction of the target material. A typical laboratory sample for macroscopic examination contains an amount of 500 grams of material. A full analysis of two replicates to assess the repeatability and one extra replicate for assessing the reproducibility would require a starting sample of 1500 grams. Samples of a smaller size would prevent to produce reasonable replicates for analysis. Four different situations can be distinguished: (A) true replicate analyses after sufficient homogenisation of large samples, (B) replicate analyses with extensive inhomogeneity in large samples, (C) a high level of inhomogeneity to be compensated by reintroduction of targeted material in small samples, and (D) lack of any duplicate analysis for the destructive nature of the method. In the presence of sufficient amounts of material per sample, the same strategy can be followed regardless of the nature of the method, either non-destructive or destructive.

The level of inhomogeneity for samples containing units in the mm size range remains much larger than the acceptable limits for relative standard deviations in analytical chemistry for comparable analyte levels (AOAC, 2002b, page 25; Codex Alimentarius, 2004, page 5; Gustavo Gonzales and Angeles Herrador, 2007), and larger than uncertainty ranges found in chemical studies in practice (De Rijk et al., 2013; Ciasca et al., 2018; Medina et al., 2018). Fluid matrices of a certain viscosity can be heated or diluted to decrease the viscosity and an acceptable homogeneity could be achieved (oil: inclusions; candy syrup: packaging material; honey: pollen). Establishing a value for reproducibility requires either a non-destructive examination procedure or options for sufficient homogenisation in the presence of a sufficient amount of sample material.

At one end of the spectrum, the analysis of subsamples of whole kernel feeds or feed materials do not represent repeatability nor reproducibility conditions, since the only relationship among laboratory samples is the situation that they are extracted from the same batch of material. In the framework of a validation study selected target material can be reintroduced in the sample matrix and re-examined after a waiting period. This approach would neglect inhomogeneity as one of the sources for uncertainty. This should be no problem as long as the information content of the validation parameter is clearly explained (Part 2, Table 3: situation C). However, the sample inhomogeneity is necessary to establish for situations in practice in order to assess the representativeness of the examined sample. Replicate examinations are needed and this approach is further discussed in section 5.4.1 (Part 2, Table 3: situation B).

The situation changes gradually when undesirable substances are smaller in size and for matrices supporting options for better homogenization. For aflatoxins, slurry mixing with water revealed a sufficient homogeneity (Reiter et al., 2009). (Semi-)liquid sample material, e.g. candy syrup or honey, or finely ground material, with grinding as part of the production process (compounds feeds) but not as sample pre-treatment, can be homogenised to a certain extent by stirring or shaking. If true replicates can be achieved after sufficient homogenisation, repeatability could be determined for comparable subsamples (Part 2, Table 3: situation A).

Validation of procedures for sample types which remain inhomogeneous and are to be examined by a destructive method where the contaminant is separated from the matrix of the sample in an automatic and/or destructive manner, would likely lack both parameters. Good argumentation of the (lack of) applicability of certain parameters needs to be included in a validation plan (Part 2, Table 3: situation D).

The different approaches for collecting data to calculate repeatability and reproducibility will be discussed in Part 2 as part of a validation plan.

4.1.6 Robustness/stability

The robustness of a method can be documented by investigating samples according to the fixed method, except for a modification of one parameter. If required, several parameters can be tested in different experiments. If recovery and bias calculated from these experiments exceed the appropriate limits, a method can still be declared validated. The conclusion is that the investigated parameters and circumstances are critical (with indication of the boundaries) for a reliable application of the method. Method parameters are usually fixed in terms of maximum time for examination and selection, time and temperature for dilution, minimum or maximum amount of material included in an analysis (sub)sample, etc.

The way in which stability is determined deviates from the normally applied procedures. The targets itself are hardly subjected to chemical decomposition and stock solutions for reference materials are not applied. In case of visual examination, environmental circumstances can influence the results of sample analysis, either in terms of number of recovered units or in terms of weight. Units such as seeds and ergot sclerotia can vary in weight due to air humidity and they can wear out. Still, the result in terms of units can be assumed to be stable under such varying circumstances.

4.2 Qualitative methods

Fragments smaller than approximately 1 mm can in most cases not be physically selected and separated. The small, microscopic size or a destructive examination prevents quantification of the units in terms of weight (mass balance, w/w). Counting is the only option. The resulting numbers are translated to compliant or non-compliant results, or to absence or presence. Low amounts can be reported as traces. If a method results in aggregation of small particles and these aggregates can be physically selected, quantitative evaluation can be applied.

The parameters common for quantitative results are not applicable for qualitative data, neither when these final qualitative conclusions are based on a translation of counts nor when achieved directly. In the absence of a final numerical result per sample, the frequency of correct or false results can be calculated from the examination of a series of identical samples with a known fixed number. This type of discrete or ordinal data follow a binomial distribution and this is a principal distinction with quantitative data, which follow a form of normal distribution with average μ and standard deviation σ. These two independent parameters relate to the two main groups of performance parameters Accuracy and Precision, respectively. A binomial distribution is described by $\mu = np$ and $\sigma = \sqrt{np(1-p)}$, with n = number of observations and p = probability of observing a positive result (Agresti et al., 2018: Chapter 6.3, page 305-315). These equations indicate a dependency of the two parameters, resulting in a relationship between the performance of a method in terms of Accuracy and Precision. This is a fundamental principle for evaluating results in a validation study of a qualitative method.

Standard descriptors for possible results of qualitative methods are based on two states: presence or absence of the target. These observations can either be correct or wrong. After combination four options are available: true negative results for blanks and true positive results for spiked samples, and either false positives and false negatives as incorrect observations for blanks or spiked samples, respectively (Table 5). This representation is a special 2x2 case of the general structure of *r x c* contingency tables (Feldsine et al., 2002; Fagerland et al., 2017: Chapter 4; Agresti et al., 2018). Generalised statistics have been developed for significance, for the influence of chance effects and for confidence limits of *r x c* contingency tables (Rand, 1971; Hubert and Arabie, 1985). In our domain, statistics specifically designed for analysing 2x2 contingency tables are commonly used. This situation renders all qualitative methods with more than two types of results (at least one additional option besides negative and positive) unsuitable for further evaluation, unless the possible results are reduced to two.

The statistical indications of Type I error (false positives) and Type II error (false negatives) can provide further statistical explanation (Sheskin, 2004, paragraph Type I and Type II errors in hypothesis testing; Sedgwick, 2014). Calculations of simulated data prove a relationship between these two types of error. If a result is based on a count, the conclusion whether a sample is free of a contaminant (negative) or is declared

to contain the contaminant (positive) can be subjected to a threshold. In general, the application of a threshold and considering results below that threshold as negative will result in all cases in higher scores for specificity and in lower scores for sensitivity [2]. The application of a threshold is primarily an issue at low contamination levels (van Raamsdonk et al., 2014). When lowering that threshold, false positive results will gradually increase in frequency and true negative decrease, with the same counts. The vice versa situation will apply to observations for spiked samples: when lowering the threshold, false-negative results will gradually decrease and true positive increase. In the view of this principal relationship between false positives versus false negatives a threshold needs a careful and precise application. A threshold should not be used to minimise either false positives or false negatives, since a decrease of one error is connected to an increase of the other. False positives can result from several sources such as lab contamination or erroneous identification. If a source has been identified, measures to avoid this can be installed. Analyses of qualitative methods which can result in more than two answers have to be reduced to two results (e.g. present/absent) for being evaluated in terms of Type I and Type II errors.

Table 5 *2x2 contingency table with frequencies of the combined output of a set of observations in a validation study or proficiency test. The addition of the contaminant, known to the organiser, is the a-priori status. The result reported by the technician or by the participants is the a-posteriori observation. Several describing statistics for column and row sums are shown. The statistics for the columns (sensitivity, specificity) and the total statistic (accuracy) are relevant for the interpretation of the results of qualitative methods.*

A-priori actual status: A-posteriori observation:	present	not_present	Row sums
present	True positive	False positive (type I error)	Positive predictive value = TP / (TP+FP)
not_present	False negative (type II error)	True negative	Negative predictive value = TN / (TN+FN)
Column sums	**Sensitivity** or true positive rate = TP / (TP+FN)	**Specificity** or true negative rate = TN / (FP+TN)	AC[1] = (TP+TN) / (TP+FP+FN+TN)

1: the term Accuracy will be replaced by a more specific term in order to respect its general use as indicator of a group of parameters opposed to Precision

A convenient way of collecting large sets of results for validation seems the organisation of collaborative studies or interlaboratory trials (ILS; ISO/IEC 17043:2010). However, prerequisites as set out for interlaboratory studies in other domains might apply to visual methods. These include sufficient training of the participants, entrance tests, test of outlying participants, and large series of samples for proper calculation of parameters (AOAC, 2002; EPA, 2016). Microbiological methods are comparable with respect to these requirements, exemplified by the interlaboratory study of two immunoassays for detection of ruminant material. These immunoassays produce a qualitative result in terms of compliance (negative) or non-compliance (positive) (van Raamsdonk et al., 2015). In the view of additional requirements for interlaboratory studies, it is recommended to produce data for validation in a set of *intra*laboratory experiments (as followed in Regulation (EU) 2021/808). At least 20 observations per treatment (spike level / matrix combination) have been recommended for microbiological validation (FDA, 2015, Table 1). This seems reasonable in the view that one false positive out of less than 20 samples would result in a score which is below the frequently used limit of 95% probability of a correct result. After establishing the correct set of validation parameters and taking care of the other already mentioned prerequisites an additional interlaboratory validation study can be organised. When using the results of an ILS with a compulsory participation, and therefore without an entrance test, two factors are included in the results: the

[2] https://en.wikipedia.org/wiki/False_positives_and_false_negatives: "When developing detection algorithms or tests, a balance must be chosen between risks of false negatives and false positives. Usually there is a threshold of how close a match to a given sample must be achieved before the algorithm reports a match. The higher this threshold, the more false negatives and the fewer false positives" (retrieved 2 June, 2021); Sheskin, 2004, page 88-89.

performance of the method and the varying performance of the participants. The latter factor should be eliminated before the results can be used to validate the method.

4.2.1 Correctness (accuracy)

In order to obtain an overall measure for the laboratory's proficiency, the number of correct positive, correct negative, false positive and false negative results can be calculated based on the produced results for each set of samples for a particular combination of contamination level and matrix type. The target values are to be set according to the composition of the test materials, either blank or spiked. The accuracy AC is the combined figure reflecting the capability of the laboratory to correctly classify both positive *and* negative samples. The statistics are to be expressed as percentages after multiplication of AC by 100. A score of 100% means that the laboratory reported correctly all positive samples as positive and all negative samples as negative. Since a score below 100% means that a part of the samples was incorrectly classified, it is of interest to know whether the reason for this result was a number of false positive and/or of false negative results. Therefore, the sensitivity and the specificity are also to be calculated, indicated as column sums in Table 5. Further background on interpretation of sensitivity and specificity will be given in the next paragraphs.

The quality parameter accuracy is an example of the group of association coefficients (Sneath and Sokal, 1972, page 129-137). These coefficients calculate in a variety of ways the association between two measures, which are an a-priori status and an a-posteriori observation in the framework of validation of methods (Table 5). ISO 13528:2015 recommends using the multistate Gower coefficient for the evaluation of qualitative results (ISO 13528:2015, paragraph 11.4.4, page 40), which is capable to handle qualitative as well as quantitative results or a mix thereof. ISO 13528:2015 does not provide guidance for the further application of this coefficient. The Simple Matching Coefficient is a derived version of Gower for two-state situations, calculated with an equation equal to that for Accuracy as shown in Table 5 (Gower, 1971; Sneath and Sokal, 1973: page 135-136). The usual term "Accuracy" for this parameter, as applied in all proficiency tests on animal proteins (see Literature, section Proficiency Tests), is confusing for its major use as indicator of a set of quality parameters, opposed to Precision (Table 3). Therefore, the name Correctness, *CS*, will be introduced, to be calculated with the same equation as applied for the former coefficient "Accuracy". ISO 13528:2015 recommends using a parameter with a comparable indication as z-scores, i.e. with optimal values close to or equal to zero. Common practice in the visual domain is to use coefficients approaching 1 or 100% to indicate the share of correct results. The ISO recommended parameters are known as distances, which can be calculated as $D_{CS} = 1 - CS$, or $D\%_{CS} = 100\% - CS\%$.

A tentative criterion could be a value for *CS* equal to or higher than 95% as indication of sufficient performance. The EURL Animal Proteins applies all results at or exceeding 90% as good or excellent scores. In simple situations, where only one type of variable is assumed, exact 95% confidence intervals of the *CS* score can be calculated from the binomial distribution. The lower limit is the smallest value of the binomial parameter P which generates the observed outcome p or higher values with at least 2.5% probability. The upper limit is the highest value of P which generates the observed outcome p or lower values with at least 2.5% probability (Macarthur and Von Holst, 2012; Wehling et al., 2011; Appendix 1).

4.2.2 Detection capability CCβ: sensitivity

The term detection capability (CCβ) originates from analytical chemistry as well, and it applies to screening methods. Definitions and procedures to calculate this parameter vary (Regulation (EU) 2021/808 Annex I paragraph 2.6; Currie, 1995 (IUPAC Recommendation); IUPAC, 2014). Most procedures for extracting this parameter assume a quantitative result. One procedure applicable to qualitative results will be discontinued by 1 January 2026 (Regulation (EU) 2021/808 Annex I paragraph 2.6.1.(b)). According to this procedure, the detection capability is defined as any level which can be detected with a significant difference from the signal of a series of blank samples. This level is then calculated from the average blank signal plus a factor times the standard deviation of that blank signal. Subsequently, the resulting value for the detection capability (CCβ) need to be tested at the level of the decision limit (CCα). However, a qualitative procedure for calculating a decision limit CCα is not included in the process for the derivation of both an average blank signal and values obtained from positive samples. Decision 2002/657/EC provides a few directions for some

parameters of qualitative methods, such as the CCβ (Annex I, paragraph 3.1.2.6). The frequency of false negatives in a range of at least 20 samples, fortified at or above a putative decision limit CCα, can be used as basis for calculating the detection capability.

Sensitivity, a much-used parameter, is calculated from samples, spiked at a certain level, and to be classified correctly as positive. This spike level cannot be justified formerly, since the CCα is not available. In practice, the detection limit CCβ can be estimated by using at least 20 samples spiked at a reasonably low level. Several spike levels can be used to find the lowest detectable level. Examples for the detection of animal proteins at low spike levels are published by Veys et al. (2010: 0.0025% successfully detected) and van Raamsdonk et al. (2014: 1 bone fragment per gram material successfully detected). In both cases the intralaboratory results were based on the examination of five samples only (homogeneity study).

Based on an assumption of β = 0.05 (5%), at least 19 out of the 20 samples to be investigated should give a correct positive result. If so, the detection limit equals the spike level used in mg/kg or weight %.

4.2.3 Selectivity/specificity

As for quantitative methods, target material should not be confused with non-target material in qualitative methods. A frequently occurring situation is the presence of fish meal in compound feed. Fish meal can be either target or non-target material. Fish meal is prohibited in compound feed for certain animals, and it can be a confusing material for the proper detection of particles originating from terrestrial animals in compound feeds where fish meal is allowed as ingredient. Staining could enhance the discrimination between target and non-target particles, but if a precise specific reaction of the staining procedure cannot be established, the method specificity will decrease.

An inclusivity and an exclusivity panel are recommended by AOAC (2016), showing the range of diversity which should successfully be identified as target material, or should be positively indicated as non-target, respectively. Such panels might be part of the design of expert systems.

Besides erroneous identification, laboratory contamination can result in false positives. Although this aspect is already mentioned for quantitative methods, the smaller particle size typical for microscopic qualitative methods would ease the presence of these particles in the laboratory environment. An emerging problem is the presence of microparticles (plastic or otherwise) in air and process water. The regular inclusion of negative control samples in sets of test samples, and the examination of traps at standard positions in the laboratory is very important (Koelmans et al., 2019).

Specificity can typically be calculated from samples with no target material. These samples may contain specified non-target material, as included in the exclusivity panel, or be blanks (only matrix material).

4.2.4 Repeatability and reproducibility: accordance and concordance

In collaborative studies based on quantitative data usually within lab repeatability and between labs reproducibility are calculated. Comparable statistics have been developed for qualitative data (e.g. microbiological detection methods) to characterise the precision of qualitative methods in a way which would be equivalent to repeatability and reproducibility (ISO, 2000; Langton et al., 2002). Accordance is the chance of finding identical results in pairs of replicates of the same treatment in the same laboratory under repeatability conditions (i.e. both found positive or both found negative). This is equivalent to repeatability for quantitative results. Concordance is the chance of finding the same result for the same treatment in two different laboratories.

Several strategies have been designed for calculating the accordance. ISO (2000) as well as Langton et al. (2002) presented calculation approaches based on the statistical model of sampling without replacement for calculating probabilities. The two values to form a pair can be sampled from the available values without replacement (Langton et al., 2002). Alternatively, the values for the replicates per treatment can be considered as representatives of a larger population of replicates and the two draws to form a pair can be sampled from that larger population, i.e. sampling with replacement is applied (ISO, 2000). Consider a bowl

with red and white units in a limited quantity. The probability to choose two white units for making a white pair is smaller when the first drawn white unit is not replaced (the set of units represents the results of a small study) compared to the situation that the first white unit is placed back (the set of units is a selection of a much larger population). This situation is comparable to the calculation of the standard deviation for a full population (denominator is n) or for a selection drawn from a larger population (denominator is n-1). The difference between the two calculation strategies decreases with an increasing number of replicates per treatment.

Practical application of the measures accordance and concordance for microscopy can be found in van Raamsdonk and van der Voet (2003), van der Voet and van Raamsdonk (2004), and van Raamsdonk et al. (2019).

4.2.5 Robustness/stability

The stability of microscopic particles or fragments can be assumed to be strong. Wearing and disturbance of the general appearance up to a certain limit can be compensated by a good identification. However, sample preparation and treatment can have a considerable effect on the detection of the particles. For example, staining of bone fragments with Alizarin Red will enhance the brittleness, which can result in more and smaller particles (van Raamsdonk et al., 2017). Certain types of microplastic show the tendency to dissolve in organic solvents or can be destructed after prolonged matrix treatment. This effect emphasizes the need for harmonisation of the procedures for sample preparation. Some reflections on these procedures or treatments are discussed in Chapter 5.1.

4.3 Estimation methods

The commonly used standards recognise only two types of analytical methods: quantitative and qualitative. Yet, there is a third type of visual method that aims at making estimations of (relative) content of an ingredient or a contaminant

4.3.1 IAG estimation model

The analysis of the botanic composition is a typical procedure for visual research. The legal basis for this examination is the obligatory label declaration of feeds, regulated for years by EU legislation and part of Regulation (EC) 767/2009. The main objective might be the transparency of trade activities, with emphasis on the prevention of economic fraud and a sufficient monitoring of feed safety. Besides the availability of a protocol, the current practices are heavily based on the existing skills of the technicians, a specific application of the general approach of expert judgment.

In order to specify a version of confidence intervals, an estimation model was developed by the IAG section Feedstuff Microscopy (method IAG-A2: Method for the Identification and Estimation of Constituents in Animal Feedingstuff). The uncertainty intervals are based on absolutely or relatively defined ranges depending on the share of an ingredient in the compound feed (Table 6). The intervals for different levels are illustrated in Figure 1. The principle for the estimation model was developed in the 1990s, and the current version was decided during the annual meeting of IAG section Feedstuff Microscopy in 2006.

Table 6 *IAG model for uncertainty analysis of the composition of a compound feed.*

Actual share in %	Accepted uncertainty limits
< 2%	"traces", not quantified
2 – 5%	+/- 100% relative
5 – 10%	+/- 5% absolute
10 – 20%	+/- 50% relative
20 – 50%	+/- 10% absolute
> 50%	+/- 20% relative

There are two aspects influencing the quality of the results of an analysis of botanic composition. The first one is the need to have supporting evidence for compensating possible under- and over-estimation. The sole visual analysis of the composition of a compound feed without complementary information lacks the possibility of confirmation and of adjustment, which might influence the precision of the final result. A way to support and, if necessary, adjust the estimated shares is to apply proximate analysis, of which Weende or proximate analysis is the classical approach (German: Weender Analyse; https://de.wikipedia.org/wiki/Futtermittelanalytik). Weende analysis, originally developed in the 19th century (Henneberg and Stohmann, 1859) provides information on basic chemical parameters: moisture, contents of ash, fat, protein and crude fibres. Since extensive information is available on the parameters of individual ingredients, the initial visually estimated shares of the several biological ingredients can be confirmed or optimized using this data. Currently several approaches exist for proximate analysis, such as Cornell Net Carbohydrate and Protein System (CNCPS), Van Soest analyses, and Near-Infrared Reflectance Spectroscopy, although these systems do not aim at the same set of parameters (Bovera et al., 2003; Godoy et al., 2016).

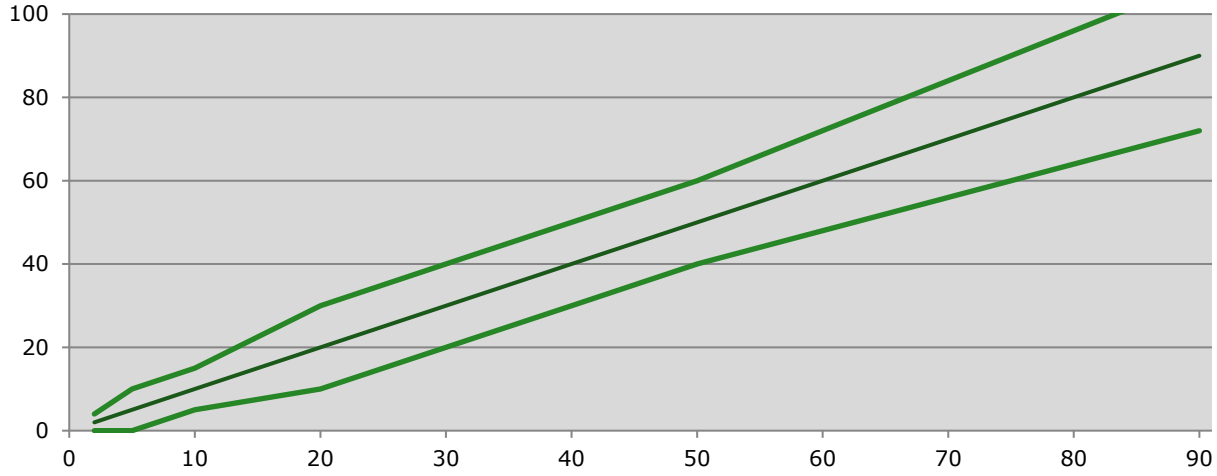


Figure 1 *IAG model for estimating uncertainty. X-axis: correct portion of ingredient in %, Y-axis: estimated portion of ingredient in %. Inner line: correct estimation, outer lines: limits for uncertainty interval at a given percentage.*

The second aspect is the ability or opportunity to detect certain materials. Some ingredients are not or hardly visible, for instance oil, molasses or very fine powders. The visibility depends on the embedding agent. Oil droplets can be made visible by Oil Red O, and spray dried milk powder will dissolve in water or chloralhydrate but remains present as particles in glycerol or paraffin oil. There is a general tendency to overestimate ingredients with a low share, and to underestimate ingredients with a higher share in the composition. If an ingredient cannot be observed, the shares of the other, visible ingredients are overestimated since the total is adjusted to 100%. Grinding will influence the recognisability of feed ingredients: fine powders are difficult to examine, if at all. Some combinations of ingredients exist which are difficult to distinguish, e.g. beet and citrus pulp (van Raamsdonk et al., 2017). These findings also illustrate the importance of a good expertise of the technician working in the visual domain. A more detailed evaluation of the relationship between method design and performance is given in Chapter 5.1.

4.3.2 Traces

The IAG estimation model includes a lower limit for estimating a quantitative indication of the share of ingredients. Below 2% the qualitative indication “traces” is used. Directive 2002/32/EC is the second framework using the term “traces”: several undesirable substances are only allowed in trace amounts which cannot be quantified. The intention is to detect these trace amounts as low as reasonably achievable, which can be assumed to be much below the 2% of the IAG estimation model. The absence of the option to

quantify these trace amounts basically imply a microscopic method. Some of the usual quality parameters such as sensitivity and specificity are relevant for the issue of detection of trace amounts.

A starting point for discussing quality parameters for the detection of trace amounts can be found in the method for detection of animal proteins in feed and feed ingredients by microscopy (Regulation (EC) 152/2009, Annex VI). The technical upper limit is a sensitivity of 0.1% as a legal demand. The results of several proficiency tests can be used to get a first impression of the lower detection limit for traces. A sensitivity of 0.96 was established for a level 0.0025% (25 ppm) in the 2009 version of the proficiency test of the EURL Animal Proteins (Veys et al., 2010). One bone fragment per 1 gram of raw sample material, with an assumed contamination level of 0.005% (50 ppm), appeared to be detectable in an IAG proficiency test. A sensitivity of SE = 0.92 was achieved by 44 participants (van Raamsdonk et al., 2014). Assuming a concentration of a factor 50 after extraction of the sediment from a compound feed with a heavy fraction of 2% (minerals), one bone fragment per 20 mg of material was successfully detected. This could be a first estimation of the sensitivity for detection of traces. A positive relationship was found between the amount of material examined and the probability to find one or more bone fragments (van Raamsdonk et al., 2014): the more material is examined the more material of the targeted undesirable material will be found. The effect of the sample size is generally recognised in statistical testing (Agresti, 2018: page 374, section "Effect of sample size").

A proper and reliable identification is another corner stone to microscopy and, hence, for the examination of traces of undesirable substances. An example is the distinction between the seed hulls of prohibited (mustard) and legalised (rapeseed) species of the genus *Brassica*. There are two conditions which need to be considered: the recognisability and manageability of the fragments, and the level of experience and skill of the technician. "Recognisability" refers to the characteristics of the fragments for identification and for distinction of fragments from mimics or look-alikes. "Manageability" refers to the ability to select and handle fragments in trace amounts. The technician needs to be trained for optimal observation and handling of the material. The paragraph on Identification support provides further information for this issue.

There is no data on the average weight of particles or fragments in a general ground compound feed, nor an estimation is available for the number of particles per milligram or gram material. This situation hampers to establish a lower limit in terms of weight percentage. Specificity and sensitivity indications in terms of frequencies can be established after examination of series of samples at low contamination levels (see Section on qualitative methods). Further research is needed.

5 Method development and QA/QC

An analytical method is composed of two basic different aspects: the procedure to handle and process the sample material for supporting proper detection of the targeted compound (the protocol), and the actual detection and identification of the target material itself (the examination). In analytical chemistry both aspects can be harmonised to a large extent. Extraction and preparation are carried out according to protocols, the parameters and operation of the equipment for detection are defined in manuals and the resulting profile is analysed with the aid of existing libraries using standard procedures. Specified procedures need to be applied in the framework of visual or microscopic observation in order to optimise the visibility of physical units. In contrast to chemical analysis, detection and identification of visible units is carried out manually, based on the expertise of the technician.

The design, development and optimisation of a method can influence the options for proper validation. A protocol should be sufficiently precise to assure transfer to and harmonisation with other laboratories. Support should be available for reliable and reproducible identification of ingredients in feed in food. The next sections will discuss types of sample preparation and factors for examination which can affect certain aspects of quality assurance.

5.1 Method design

In order to assure an application of a method in a laboratory, matching the requirements of the main Quality demands Precision and Accuracy, and to achieve a sufficient transferability of a method to other laboratories, protocols as precise as possible should be available for proper harmonisation. In this section several elements for material treatment, and for detection and identification will be discussed for their relationship with parameters of quality assurance. There are two main aspects in this matter: (a) certain treatments can influence the number of particles, such as sieving and grinding, and (b) certain treatments can influence the recognisability of the particles, such as staining reactions and the use of specific embedding agents.

5.1.1 Material treatment and work flow

The efficiency of the search for visible units can be optimised by three different types of actions. The contaminant or specific ingredient can be concentrated (extraction), it can be given a distinguishing colour (staining), and fractions of sample material with comparable particle size can be separated (sieving). Several of these types of actions can be combined. The way these actions can be applied depends on a case-by-case approach. Another aspect is the need to reduce the sample to one or more subsamples, still assuring sufficient homogeneity.

Extraction
The sensitivity of a method can be increased if a treatment would be applicable for concentration or extraction of the target. Particularly contaminants with a particle size below 1 mm cannot be concentrated by physical selection using a pair of tweezers. Concentration might be achievable based on their specific density. A frequently applied situation is the extraction of the heavy fraction from compound feeds, which might contain bone fragments of PAPs (Regulation (EC) 152/2009, Annex VI). Another application is the separation of microplastic particles (Pagter et al., 2018). Centrifugation of fluid matrices for separation of solid fraction can be applied in several occasions, e.g. pollen in diluted honey, pellet material in warm oil (Ohe et al., 2004; Regulation (EC) 152/2009, Annex VI, non-consolidated version). Packaging material in diluted confectionary syrup can be extracted by sieving.
Effect: higher sensitivity.

Staining
Differential staining is a technique which is intended to achieve a better visibility or detection of the target material. A staining technique should precisely stain the particles of the target material and leave the other particles unstained. Bone particles can be differentially stained by using Alizarin Red. However, this dye can stain specific minerals as well, and fragmentation can occur as process by-effect (van Raamsdonk et al., 2017). Other examples are Chlorazol Black for chitin parts of insects, lugol staining for starch or for muscle fibres, cysteine reagent for keratin material (hair, feather), Tetramethylbenzidine for blood plasma, Nile Red or iDye stains for microplastic particles (Ottoboni et al., 2017; van Raamsdonk et al., 2012a; van Raamsdonk et al., 2011, Maes et al., 2017; Karakolis et al., 2019). In specific cases a contaminant has naturally a distinguishing colour, e.g. ergot sclerotia (black) in light coloured grains of cereals (van Raamsdonk et al., 2016: front cover).
Effect: higher specificity depending on the specific colouring reaction of the dye, deviating recovery due to increase (fragmentation) of the number of particles in the particular case of Alizarin staining.

Sieving
The separation of several sieve fractions containing material of different particle size (coarse, fine) will enhance the detection of a specific ingredient or contaminant: the attention of the technician is not distracted by (large) differences in particle size. From a technical perspective sieving might be necessary for a proper application of the equipment. Fine fractions are primarily used for the preparation of microscopic slides to be examined at higher magnification (compound microscope), and coarse fractions are intended for observation at low magnification (stereo microscope) or with the naked eye. Examples of current methods are the detection of animal proteins in feed, determination of Ambrosia seeds in bird feed, and packaging material in bakery by-products (Regulation (EC) 152/2009 Annex VI; IAG method A5; van Raamsdonk et al., 2012c).
Effect: either higher sensitivity and/or specificity. Sieves need to be thoroughly cleaned in order to avoid lab contamination, which would result in false positives (lower specificity).

Homogeneity
The use of a subsample in feed analysis is allowed for microscopic procedures according to the Regulation (EC) 152/2009 Annex II. This subsample should be representative for the original laboratory sample. Grinding can be used for homogenisation of sample material. This has an effect on the particle size and, hence, will result in larger numbers of smaller particles. In those cases where particles need to be counted, additional grinding for the sole reason of homogenisation should be avoided (animal proteins: Regulation (EC) 152/2009 Annex VI; van Raamsdonk et al., 2014). Also, smaller particles might be more difficult to detect and identify. In a range of other cases grinding cannot be applied, such as seeds in bird feed, ergot sclerotia in whole grain cereals, packaging material in former food products. Solutions are the examination of the entire laboratory sample, or using a splitter for the production of subsamples that can then be investigated separately. Even then, the resulting subsamples can differ substantially.

Grinding
An example for careful harmonization is the application of grinding. In feed analysis, grinding is often used primarily to separate particles and fragments that have been extruded to create a pelleted product. Nevertheless, the final target of milling or grinding is size reduction of particles. In the assumption that the amount of material will remain the same before and after milling (besides the accidental loss due to the processing), size reduction will automatically result in more particles. In a general sense, size reduction by wearing or abrasion such as found in plastic will result in smaller but more numerous particles (van Raamsdonk et al., 2020). There is a large range of different types of mills (Berk, 2018), many of them for industrial application. In all cases the same principle target applies: size reduction of particles. For laboratory purposes primarily rotor mills (cassette or hammer mills, examples of the group of impact mills) and knife mills (or household coffee grinder, examples of attrition mills) are in use. For all types at least one parameter in the process will correlate with the size reduction. Several studies provide information on Particle Size Distribution (PSD) after grinding with some type of impact mill (Indira et al., 2006; Shashidar et al., 2013; Steglich et al., 2015; Naimi et al., 2016). In all these studies several mesh sizes of the sieve in a rotor mill were used to present the different PSDs. One study was found using a type of knife mill (Lee et al., 2014). Here the development of the PSD was correlated with time (Figure 3 in Lee et al., 2014). After 30-60 seconds a steady state was reached. The coarser fractions reached a stable volume at an earlier time compared to the fractions

with smaller particles sizes (Figure 5 in Lee et al., 2014). In the current version of the method for detection of PAPs in compound feeds a knife mill is allowed to be used without a fixed duration (Veys et al., 2017), although different durations for milling would result in different PSDs. A knife mill should be avoided for lack of the opportunity of harmonisation in the view that time in terms of seconds is difficult to control. Alternatively, the effect of grinding with a rotor mill on the PSD is relatively easy to monitor, which consequently results from several studies, including Veys et al. (2017). As for sieves, grinding equipment need to be thoroughly cleaned in order to avoid cross-contamination, which would result in false positives (lower specificity).
Effect: an increased number of particles will have a larger probability to exceed a threshold (biased sensitivity), smaller particles are more difficult to recognise (lower specificity).

All mentioned procedures for sample treatment can and should be harmonized to a large extent. If certain types of matrices make it necessary to apply dedicated and deviating procedures, specific exceptions can be made in the framework of a method. Such situations should be precisely described and rarely chosen. The choice to install modifications for a mainstream type of matrix should strongly be avoided. In those circumstances the scope of a particular method should be narrowed, and a comparable but slightly different method should be designed for every frequently occurring type of sample or material. An example is the procedural difference between matrices with low fat content versus those with high fat content or consisting exclusively of fat or oil. This strategy would allow to establish conformity and equivalency of methods, which is addressed in a next chapter.

5.1.2 Examination

The final step of microscopic evaluation is not carried out by an instrument (e.g. mass spectroscope, near-infrared spectroscope, PCR thermocycler, Biacore reader), but is carried out by the technician. The implication is that the protocol for a microscopic method should be harmonised as far as possible, up to and including grinding, sedimentation, sieving, staining, amount of material examined, number of repetitions and number of slides. The elements influencing proper examination by the technician should match the experience of the technician. These circumstances include the embedding agent, in certain applications containing a staining component, the choice of the aperture of the condenser, the use of polarisation or filters, among other factors. A major part influencing the appearance of particles during microscopic evaluation is the embedding agent, either with or without dying properties. This element of several methods will be discussed as example of the relationship between a methodological element and the individual requirements of the technician.

Data from proficiency testing for the detection of animal materials in compound feed revealed that embedding agents with several different viscosities are frequently used (see surveys in annual reports of IAG ring tests; van Raamsdonk et al., 2019). Differences in performance of the method after using an embedding agent with high viscosity versus low viscosity were never found. Viscosity is a delicate aspect for microscopic identification. It is the property of a fluid to resist to sheer deformation, due to cohesion between the molecules. Fluids with a high viscosity deform more slowly than fluids with a low viscosity. Fluids with a higher viscosity will be slower to enter porous particles in a slide, which results in air-filled (black) holes in the particles. Fluids with a too high viscosity take a prolonged time to enter the mass of a porous particle. It is especially of importance for the classification of bone particles because: a) osteocytes in bone particles act as air-filled holes and b) the shape and distribution pattern of the osteocytes, as well as the visibility of the canaliculi, allow the distinction between classes of vertebrates (fish versus terrestrial vertebrates). Figure 2 shows the effect of different viscosities in the low and mid-range values (Source: project STRATFEED, ARIES version 1.0, 2004). In the STRATFEED project glycerol was not considered. At higher viscosity, the general pattern of distribution of the osteocytes is clearer, but at lower viscosity the shape of the osteocytes and the canaliculi will be more visible (at higher magnification and taking advantage of the absence of an air bubble).

Since recognition of particles in microscopy is primarily a matter of training and experience, the shift from using an embedding agent used in the training phase of a technician to an agent with a fully different viscosity needs additional training and documentation. Besides the factor of a different appearance of bone particles, the major issue is the match between a particular appearance and the experience of a technician. This aspect was never investigated.

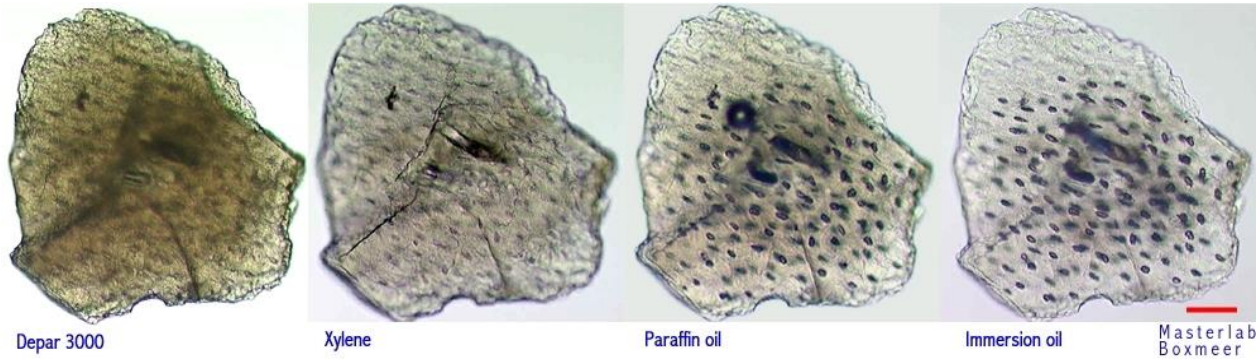


Figure 2 *One bone particle embedded in four different agents. The particle was extensively rinsed between the different applications. Order of embedding agents applied was from low to high viscosity. Source: ARIES version 1.0, 2004; Masterlab, Boxmeer, the Netherlands. Scale: 50 µm.*

Paraffin oil and immersion oil were suspected to have a variable composition with the consequence of a presumed uncontrolled behaviour. However, viscosity seems to be a parameter with a certain fluctuation anyway. Norland provides different values for the viscosity of the allowed Norland adhesive 65: 1000 cps at https://www.norlandprod.com/adhchart.html, 1200 cps at https://www.norlandprod.com/adhesives/NOA%2065.html (retrieved 27 September, 2022).

Embedding agents, being fluids, may have dissolving properties. Particles could be seen in one embedding agent and be dissolved in another (spray dried milk powder in lye, starch globules, sugar crystals or salts in water, oil or fat globules in apolar agents). Alternatively, particles composed of cells and membranes could swell or shrink depending on the salinity of the embedding agent (osmosis).

The reports of the annual IAG proficiency tests include a survey of method parameters, the use of embedding agents among them. The most frequently used embedding agents were and still are glycerol and paraffin oil. The latter is available in several versions with different viscosities. The version with a viscosity of 110-230 mPa·s is most commonly used. Immersion oil has a viscosity comparable to that of paraffin oil. An evaluation of the participants' performance using the four mentioned embedding agents is possible based on the publicly present data in the Annual reports of the IAG proficiency tests. The results in the years 2015-2019 were evaluated considering the results below the decision limit as negative (official EURL approach). Data have been collected for four specific combinations: PAPs in the presence of fish (0.1% PAP with 0.1-2% fish), PAPs in the absence of fish (0.01%-0.1% PAP), absence of PAPs and presence of fish (0.1%-2% fish), and blanks. Not all combinations had been included in every targeted year. The scores for the presence of PAPs and fish were collected for the four combinations, resulting in a maximum of eight data points per participant per year. One participant produced three errors in two subsequent years, and the results of this participant were excluded for those respective years. The general performance is CS=0.95 for paraffin oil (n=374 datapoints) and CS=0.96 for glycerol (n=508). Other data is collected in Table 7.

Table 7 *Describing parameters (refractive index and viscosity), performance (overall correctness), number of datapoints available and participation, extracted from data of five IAG proficiency tests (2015-2019).*

Embedding agent	Refractive index	viscosity (mPa·s)	Correctness	Datapoints	N participants/yr
Glycerol	1.47	1490	0.96	508	12-17
Paraffin oil	1.48	110-230	0.95	374	12-13
Immersion oil	1.51	100-120	0.95	274	7-10
Norland adhesive 65	1.52	1000 or 1200	0.98	130	3-6

Sources: correct detection of either PAP or fish in four combinations: blank, fish only, PAP only, fish+PAP.

Results used of samples: 2015: B, C, D; 2016: A, B; 2017: B, C, D; 2018: A, B, C, D; 2019: A, B, C, D.

The overall performance of all four embedding agents is excellent according to the criteria of the EURL (Fumiere et al., 2021: consolidated CS, expressed as AC, at or exceeding 0.90). In terms of refractive index

both immersion oil and NOA65 have an almost perfect match with that of a standard cover glass. In the view of the performance of paraffin oil and the possibility to define this embedding agent in the same way as NOA 65 for their varying viscosities, viscous paraffin oil should be accepted as option in the method for the detection of animal proteins. In the view of its performance and almost perfect refractive index, immersion oil should be accepted as well. The EURL confirmed the legitimate use of paraffin oil and immersion oil within the requirements of the official method by mail in August 2021 (Regulation (EC) 152/2009, amended by Regulation (EU) 2020/1560, Annex VI, paragraph 2.1.2.1.3.2).

5.2 Identification

A proper and reliable identification of the target, either a contaminant, an undesirable substance or a legal ingredient, is for most visual methods based on expert judgment. A confirmed identity or classification is the basic requirement for establishing the specificity of a method, and for assuring a good detection of targets during the daily workflow of monitoring, providing that other sources for false positives such as laboratory contamination are absent. The establishment of the identity of or a classification for a material is the basis for solving authenticity issues. This might cover a large range of different products and materials. In all cases a reference should be available for providing the correct identity of the target and confirming the established result.

The visual identification of specimen is a very common practice in zoology, botany and mineralogy, with a large historic track record and extensive documentation. In botany not only plants but plant parts such as pollen or seeds are treated as separate entities, either fresh or dried as herbarium specimen, and relatively well documented. Even then, expert views can differ on the identity of a specimen. Some legally addressed undesirable components (Directive (EC) 2002/32: whole seeds, ergot sclerotia; Regulation (EC) 152/2009: Annex VI: processed animal proteins) are well documented. For other categories of materials, such as processed by-products of oil or starch production, herbs, spices, or inorganic material (packaging material) documentation is specialised, scarce or absent. Still, justification of an identification is important.

The way such confirmation can be achieved depends on the aim of the method. Identification of feed ingredients for label control or composition analysis, or pollen profiling of honey would hardly allow analysis of duplicate samples for reasons of cost effectiveness. In the case of identification of microplastic, which is usually achieved by applying an Fourier Transformation InfraRed microscope and libraries of FTIR profiles, duplicate analyses can be carried out more easily (Käppler et al., 2016). This step in the procedure for microplastic detection is necessary in order to avoid detection of particles or fibres which are not considered microplastic or are otherwise excluded from the scope of the method.

Inclusivity and exclusivity panels should support proper discrimination. Duplicate identification by a second technician or support by means of a Decision Support System might be feasible or considered necessary for safety enforcement.

5.2.1 Identification support

A way to confirm an identification is by means of expert judgment. Two experts are usually not available in one laboratory. Documentation for proper recognition of particles or fragments is available. Historical descriptions, images, drawings and tables with discriminating features are published in hand or reference books. Some frequently used works are:

a. General overviews of food and feed materials with descriptions of characteristics: Gassner, 1973; Mészáros and Bihler, 1983; Gassner et al., 1989; Hahn and Michaelsen, 1996; Hohmann, 2006.
b. Overviews focusing on herbs and spices: Fischer and Kartnig, 1978; Eschrich, 1999; Hohmann et al., 2001; Rahfeld, 2009; saffron: Alonso et al., 1998.
c. Identification of starch: Seidemann, 1966; Czaja, 1969.
d. Pollen identification: Sawyer, 1981, 1988; Persano Oddo and Piro, 2004.
e. Seed Atlas: Cappers et al., 2012.
f. Methodology: Feigl, 1958; Flint, 1994; Gassner et al., 1989; Hohmann et al., 2001; Hohmann, 2006; Sawyer, 1981, 1988.
g. Combination of chemical and microscopic descriptions: Winton and Barber Winton, 1939, 1946.

Physical collections of reference materials are also important for extension of written documentation. Examination of reference materials allows to use preferred embedding agents, specific treatment or staining and three-dimensional observations. Proper documentation and labelling should be fully assured in order to make correct observations and comparison with the subject material. Collections can be maintained by own management (e.g. seeds, feed ingredients, slide collections of animal particles or pollen) or by other official institutes (herbaria, forensic laboratories).

Knowledge or expert systems can provide interactive support for the process of identification. The class of knowledge systems applied in the framework of feed and food quality and safety can be indicated as decision support systems or classification systems (Determinator: van Raamsdonk et al., 2012d). Some further information on the value of expert systems and way of application for visual research is provided in Appendix 3.

5.2.2 Statistical tests

The reproducibility of a previous classification or identification of samples can be confirmed by the application of statistical tests. The results of qualitative methods in terms of the pairs of identifications per sample (either a correctly or wrongly reproduced result) can be evaluated by calculating concordance or other statistics, such as Cohen's Kappa (Stehman, 1997; Langton et al., 2002; McHugh, 2012). For quantitative results the Mann-Whitney U test (non-parametric, independent samples), Wilcoxon signed rank test (non-parametric, matched results) or paired or unpaired t-tests (parametric) can be applied (Fay and Proschan, 2010). The Kruskal-Wallis H-test is a non-parametric extension to these tests, which can be used for more than two groups of equal or different sizes. An overview of parametric and non-parametric tests can be found in Sheskin (2004).

5.2.3 Probability of Identification

A procedure for establishing a Probability of Identification (POI) has been developed (AOAC, 2012). The principle is to establish the level at which a "botanical" or an undesirable substance can be identified with a 95% confidence. The example used in the report is the identification of American Ginseng mixed in Asian Ginseng as matrix. Both inclusivity and exclusivity panels are provided. The detection is based on a chemical marker.

The principle of POI is developed from the measure Probability of Detection (POD), as is stated in the AOAC report. In this way the POI is just a special case of developing a level of detection. Correct identification (in terms of a Boolean) is documented by the panels for inclusivity (correct: yes) and for exclusivity (correct: no). Identification of biological entities is depending on the appropriate classification system, and this is not documented by the POI.

5.2.4 Strategies for validation of identification procedures

An approval of the process for establishing the identity of a sample can be reached by repeated analysis of known samples. The documentation of the validity of the process should contain information on, in an ideal situation, both Accuracy and Precision. Besides justification of the contents of an expert system, the same procedure can be used to validate the expertise level of experts.

There are two options:

- **The identity of the sample(s) is known and approved**. This means basically that reference material will be used. This reference material should have a traceable and documented source (PAPs of known and uncontaminated source, pollen and seeds collected from the original plant, feed ingredients from approved manufacturers). Material from existing collections could be valuable to use. Approval of identity by means of identification based on a non-visual source (chemical marker, DNA) could help to establish correct a-priori documentation. Results are a documentation for both Accuracy and Precision in all cases where the second (blind, a-posteriori) identification matches the a-priori identity.
- **The identity of the sample(s) is not known**. The first and second analysis have the same reliability, and information in terms of a-priori known identity does not exist. The documentation will only support

precision, since theoretically a wrong identity can be achieved twice. However, provided a sufficient expertise level of the technician(s), the chance of drawing the same wrong conclusion twice is the square of the chance of drawing the same wrong conclusion once. There are several options for calculating the relevant statistics. These include concordance (within-lab reproducibility) and Cohen's Kappa statistic for qualitative methods (ordinal results), and Mann-Whitney U test, Wilcoxon rank order test or paired or unpaired t-tests for quantitative results.

Other choices to be made are the type of classification system: multi-class or two-class. The latter will result in the conclusion: the investigated material belongs to type A or not to type A (A or ¬A). Finally, validation can be intended to focus at the final identity, which would need a multi-class system with a large array of classes, or at the level of single characteristics. The latter eases the process, since expression in terms of present/absent, yes/no, etc. (yes or ¬yes (no)) will apply for a range of characteristics.

5.3 Conformity and equivalency of methods

Once a method has been developed and distributed among laboratories, harmonisation of the implementations should be achieved. There are several levels of harmonisation, and the application of requirements should be defined.

Conformity is defined as a precise implementation of a method (the protocol) without deviations.

Equivalency means that a method is implemented including (minor) modifications. These modifications can include an extended scope, a larger range of contamination levels or treatments of samples different than included in the original method. Modified elements of treatments need to be validated or approved before implementation. Qualitative methods with an identical probability of detection under the same conditions might be considered equivalent (Macarthur and von Holst, 2012). Otherwise, a modified version with a deviating performance should be indicated as a laboratory version of an official method.

The requirements of conformity or, alternatively, of the equivalency need to be specified for the two different aspects of visual methods: material treatment (the protocol), and detection and identification (the examination). Harmonisation needs to be assured for procedures and treatments such as sieving, staining and grinding. Aspects relevant for detection and identification, those elements with a relationship to the specific expertise of the technician, should not be intended to be harmonised among the laboratories. Instead, conformity should be achieved among the different methods with comparable types of observations, as practiced by the same (group of) technicians. This leaves room for maximising the usability of the specific expertise, provided that the validation by an in-house procedure of the local implementation of an official or other method meets the required probability of detection. Subsequently an interlaboratory validation study can be organised providing the a-priori existence of a documented conformity of the local implementations and a sufficient mastering of the entire method by all participants.

5.4 Quality control of sample analyses

The application of a validated method is no guarantee for correct results at all times and quality control measures need to be installed. Several measures can be taken to document and assure the necessary reliability of test results. The two elements of a visual method should be distinguished: the workflow of the protocol and the procedure for examination.

5.4.1 Workflow

The daily practice of the application of a method should be controlled in various ways. Several levels of quality control are defined: duplicate analysis, control samples and participation in interlaboratory studies or proficiency tests.

As argued, duplicate analyses (**level 1**) are impossible or unreasonable for most methods of visual inspection. A solution can be to examine a sample twice by two technicians, or with the aid of an expert system to check an initial conclusion. The chance of drawing a wrong conclusion twice is the square of drawing the same wrong conclusion once, which might serve as indication of a certain level of reliability of the final result. A selection of samples can be stored for a second opinion in order to confirm the initially made observations during a fixed period of time, e.g. a month.

A procedure has been developed for the examination of undesirable substances in whole kernel materials which includes the calculation of a relative standard deviation as parameter for quality control (Figure 3).This procedure for quantitative methods with a maximum legal limit (weed seeds in bird feed, ergot sclerotia or Besatz in cereals) consists of the division of a sample in four subsamples of equal weight, preferably of approximately 500 grams each, with a deviation of +/- 5 gram or an SD below 10 g which is below a deviation of 2% (Peereboom et al., 2021). Two of these subsamples will be analysed and the relative standard deviation will be calculated for process control. In those cases that the level of contamination in these two subsamples, calculated as weighted mean of the results of the two subsamples, does not exceed the analytical threshold of the undesirable substance(s) found, the total amount will be reported in terms of mg/kg or % (w/w). If the amount found is exceeding this analytical threshold, the second set of two subsamples will be analysed as well, and the total amount found in the total of approx. 2 kg sample material will be reported as mg/kg or % (w/w), calculated as weighted mean of the results of all four subsamples. For reasons of verification, the analysis of the second set of two subsamples is recommended to be analysed for one in ten samples with a mass fraction of the undesirable substance between 50% and 100% of the respective analytical threshold. As indicated in the flow chart of Figure 3, an intermediate result could be stored in order to monitor occasional deviations.

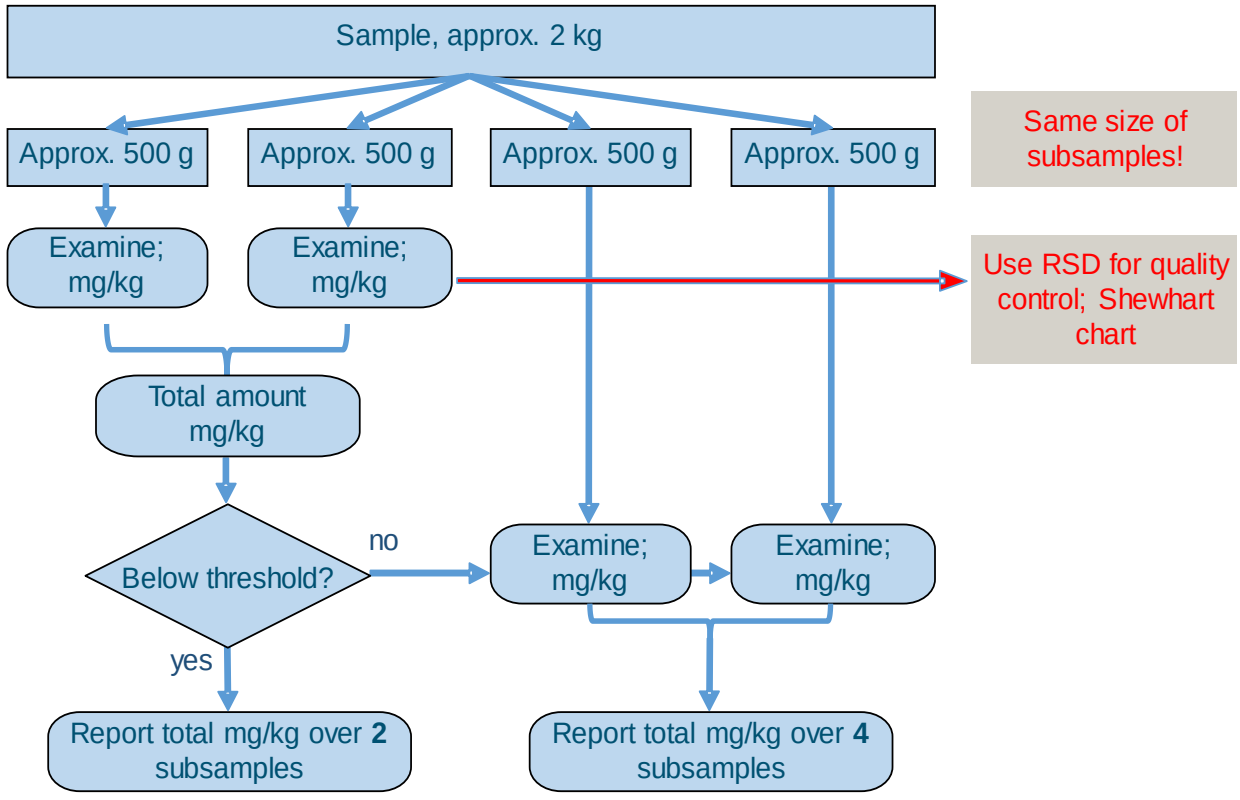


Figure 3 *Flow chart for the detection of undesirable substances in whole kernel sample material from practice (seeds, sclerotia).*

The method as presented in Figure 3 is the situation matching Regulation (EU) No 691/2013, amending Regulation (EC) 152/2009 in terms of the amount of material to be examined. In the process of revision of Regulation (EC) 152/2009 with the intention to compile a fully revised version by the year 2022, a simplified version was agreed. In this version, one portion will be examined and in the situation of exceedance of a

threshold a second portion will be examined. Threshold values will be lower according to smaller amounts of examined material (van Raamsdonk and van der Voet, 2022). The appropriate threshold levels for five combinations of undesired substances and legal limits are presented in Table 8.

The initial analysis of one or of two portions is not comparable to a screening method, nor is the extended analysis of one or two additional portion(s) comparable to confirmation. This is illustrated by the different applicable quality parameters (see Table 3), and the situation that the result from screening is not part of the final result.

Table 8 *Estimated thresholds for screening results (one or two portions). Data extracted from van Raamsdonk and van der Voet, 2022.*

Target	Legal limit (mg/kg)	Threshold (mg/kg) screen 1 of 4 portions	Threshold (% of legal limit) screen 1 of 4 portions	Threshold (mg/kg) screen 2 of 4 portions	Threshold (% of legal limit) screen 2 of 4 portions
Ambrosia seeds	50	17.5	35%	33.2	66%
Datura seeds	1000	857	86%	907	91%
Ergot sclerotia, food	200	0	0%	70	35%
Ergot sclerotia, feed	500	230	46%	312	62%
Ergot sclerotia, feed	1000	642	64%	726	73%

Control samples play an important role in the range of options for process monitoring (**level 2**). The analyses of control samples at regular time intervals provide the possibility to assure absence of laboratory contamination and a repeating assurance of correct identification. Control samples of a range of types are defined for molecular biological tests (ISO, 2006, 2019). Special attention needs to be given to contaminants which are or can be regularly present in media (microbiology), in matrix material as background contamination (chemical process contaminants: dioxins), or in fluids and/or air (hairs, fibres from clothing, microplastic). Especially the background presence of anthropogenic particles or fibres in water or air, or separated from plastic disposals, influencing microplastic detection is a major source of deviating results (Koelmans et al., 2019). Control or Shewhart charts allow to control intermediate measures or results of methods. Criteria for non-performance and correcting actions can be found in e.g. Wheeler (2000).

In the range of options for control samples, the most appropriate ones for visual methods are:

1. Negative controls: test samples consisting of the most appropriate matrix without the target under study. If the samples are all based on the same composition, the deviations through time can be measured. Control or Shewhart charts can be used for monitoring a quantitative value.
   Purpose: assessing laboratory contamination, proper workflow. Possible frequency: one in every fifty samples, weekly, or as part of every batch of samples.
2. Positive controls: test samples consisting of the most appropriate matrix containing the most appropriate target(s) of the inclusivity panel. The spike level should be below the legal limit, if any, for that contaminant or undesirable substance. Correct results for samples with a spike level exceeding the legal limit still does not prove the capability of the technician or laboratory to perform correctly at that legal level.
   Purpose: assessing proper identification, specificity. Possible frequency: biannual or quarterly. The use of at least two different stocks for producing the positive controls with different levels of contamination, used intermittently, is recommended.
3. Environmental controls: traps at specified locations in the laboratory for collecting airborne fragments of any kind.
   Purpose: assessing laboratory contamination. Possible frequency: biannual or quarterly examination.

Especially the control samples of types 1) and 3) are of major importance for quality assurance in the detection and identification of microplastic. Fluids, vials, equipment, lab coats and air need to be tested on the presence or emission of microparticles.

Some visual methods include the evaluation of procedural measures, which can be used to monitor the correct application of a method, or of a certain action within a method. An example is the monitoring of the tetrachloroethylene (TCE) sediment amount of a compound feed used as control sample which is examined at regular intervals. This measure can serve three goals: 1) checking the sedimentation (did the separation with TCE work well); 2) checking the homogeneity between subsamples (if a sample is kept in a jar for some time, the heavy particles might sink to the bottom of the jar with time; careful mixing of the sample before taking a subsample for analysis is needed); 3) the amount of sediment can give an idea of the effectiveness of the analysis when examining only a limited number of slides (a given number of bone particles will be more diluted in a large sediment). The summed weights of sediment fractions or sieve fractions should have the same amount as the total weight of the initial sample, within reasonable limits. Some circumstances might prevent a weight balance check, for instance when the fat component was dissolved in an organic solvent.

The participation in interlaboratory studies (ILSs) or proficiency tests (PTs) is an important part in quality control (**level 3**). The organisation of reliable ILSs or PTs is an elaborate process which deserves a formal platform and organisation structure. ISO 17043:2010 provides a set of requirements for organizing ILSs and for the evaluation of the results. Procedural factors such as communication with participants, design, operation of testing schemes, reporting, confidentiality and management requirements are included in the Standard. These protocols can be applied in general to proficiency testing. The part on evaluation of the results needs special attention. Section 4.7.2 of ISO 17043:2010 indicate the existence of "situations where unusual factors make evaluation of results and commentary on performance impossible". Such a situation applies to visual methods.

Annex B of ISO 17043:2010, indicated as "informative", provides procedures for evaluation of results:

A. Quantitative results: measurement uncertainty, testing on outliers or the application of robust statistics, and the calculation of z-scores are mentioned. These statistics are only valid if the detection method under study is eligible to calculate such statistics. Measurement uncertainty has been addressed in a previous section of this Guidance. Z-scores are calculated from the average recovery and the standard deviation of the reported results of the participants, under the assumption that all samples as analysed by the participants were taken from a homogenised batch, i.e. with exactly the same level of contamination. The reported results are assumed as draws from the same distribution, usually a standard normal distribution (ISO 13528:2015: page 7). ILSs of quantitative visual methods are mostly based on individual spiking of every sample in order to avoid problems of inhomogeneity. This means that every sample is the only representative of an independent population, and only one result from a population of possible results is available. An assigned value *X* of the original batch or a standard deviation σ of the reported results does not exist, preventing the calculation of Z-scores.
   ISO 17043:2010 Annex B and ISO 13528:2015 propose several alternatives for Z-scores, of which the deviation of the reported result from the assigned value is applicable: every individual sample has its own assigned value, and for every sample a reported value exists. The calculation of the performance preferably follows two different approaches, depending on the type of result:
   a. Number or count: equation $D = (x - X)$, with x as reported value and X as assigned value (ISO 17043:2010, Annex B.3.1.3 a; ISO 13528:2015, paragraph 9.3.1).
   b. Level of contamination (w/w) expressed as percentage: equation $D_{\%} = \frac{(x-X)}{X} * 100\%$, with x as reported value and X as assigned value (ISO 17043:2010, Annex B.3.1.3 b; ISO13528:2015, paragraph 9.3.1). This value relates to recovery as $D\% = R\% - 100\%$.

   The interpretation of the difference between the reported result and the assigned value cannot follow the criteria as expressed in ISO 13528:2015 section 7.2.2, since the standard deviation does not exist. Criteria for interpreting performance in interlaboratory studies of quantitative methods is poorly documented. The recovery intervals as shown in Table 4 apply to situations where a main sample is divided in two or more portions. A well-trained technician can distinguish the target from the matrix and from mimicking material with high precision. Occasionally, a (small) seed or small fragment of an ergot sclerotium can be overlooked, and a fragment can fall apart in two fragments. Reasonable limits for counts might be +/- 1 seeds or +/- 2 fragments of ergot sclerotia. Limits of +/- 5% for the recovered level of contamination would be an analogy to a 95% confidence interval. The PT on ergot sclerotia in rye grains showed acceptable results when applying such limits (Peereboom et al., 2021).

B. Qualitative and ordinal results: ISO 17043:2010 Annex B section 3.2 states that the usual statistics are not appropriate to be applied, leaving room for alternative approaches such as ranking or indication of percentiles. ISO 13528: 2015 (Chapter 11) proposes to use the Gower coefficient. This is a multistate parameter, which can be modified to a two-state coefficient. Some more details are given in the paragraph Qualitative methods, Correctness (4.2.1). Calculation of the combined parameter Correctness (formerly known as Accuracy) along with Sensitivity and Specificity (Table 5) and the calculation of Concordance and Accordance is applicable as shown in a series of PT reports (see relevant section in Literature).

5.4.2 Evaluation and documentation of the results

The principle that proper observation of particles is dependent on the skills and expertise of the technician applies to the daily work of sample examination as well. In principle, every identification of an individual unit should be based on a documented library of possible identifications for later control. This situation is comparable to the use of libraries with DNA sequences or MS signals for identification of results in molecular biology or analytical chemistry, respectively. Although a lot of documentation is published in handbooks for a variety of types of target material, the interactive expert system Determinator provide the option to store and retrieve the choices made to arrive at a specific result of identification sessions (see paragraph Identification). The records of an identification stored in Determinator can be exchanged for contra-expertise or second opinion when this expert system finds a wide application among laboratories. Another advantage of expert systems in daily routine is the use as tool for training. The stored background documentation of the results as achieved by the trainee can be evaluated by the trainer and used for improvement. Several datamodels are being published or in final stages of development, including animal by-products (focusing on sediment inclusions), feed ingredients, starch, pollen, ragwort and mimicking plants, *Ambrosia* seeds, *Brassica* seeds, rice classification for custom purposes.

The result of the examination should be properly documented. In contrast to other methods in other domains, there is no record in terms of a profile or a sequence. Documentation should be stored in versions that provide information on the examination process as followed by the technician. Materials that has been weighted are physically present. Basically, this applies to all quantitative methods: weed seeds, ergot sclerotia, packaging material, Besatz. The selected portion can be stored for later reference for a period of e.g. five years. The material that is observed and identified in qualitative methods is usually smaller than 1 mm (animal proteins, weed seed fragments, microplastic, pollen) and is usually not physically separated from the matrix. Physical storage can be achieved by making permanent slides using Norland adhesive or another permanent embedding agent. “Glass archives” can be informative as reference material or for training.

Documentation for future reference can be collected as images as well. Photographs should preferably represent the view of the material in the slide. For specific substances (e.g. bone fragments, pollen grains) several, at least three, focus layers should be photographed in order to allow to “scroll” along the depth of the particles. In case of identity or composition (ingredients in feed, starch grain mixtures, packaging materials, pollen in honey) both images at low and high magnification are necessary in order to show both the overview of the sample material and the details of individual units. When relevant, additional images with polarised light should be made, for instance for starch grains. A size bar should be included. Names of image files should indicate the sample number, a rank number, magnification, embedding agent, staining type and/or illumination (polarization, fluorescence) for allowing easy access to the image library. A logic organisation of images in a structure of folders and subfolders is recommended.

Literature

General

Agresti, A., C. Franklin, B. Klingenberg, M. Posner, 2018. Statistics. The Art and Science of Learning from Data. Pearson Education Ltd., Harlow, England.

Amato, G., R. Desiato, T. Giovannini, L. Pinotti, M. Tretola, M. Gili & D. Marchis, 2017. Gravimetric quantitative determination of packaging residues in feed from former food, Food Additives & Contaminants: Part A, 34: 1446-1450.

AOAC, 2002a. Appendix D: Guidelines for collaborative study procedures to validate characteristics of a method of analysis. http://www.eoma.aoac.org/appendices.asp

AOAC, 2002b. Appendix K: Guidelines for Single Laboratory Validation of Chemical Methods for Dietary Supplements and Botanicals. http://www.eoma.aoac.org/appendices.asp

AOAC, 2016. Appendix F: Guidelines for Standard Method Performance Requirements. http://www.eoma.aoac.org/appendices.asp

Berk, Z., 2018. Food process engineering and technology, third edition. Chapter 6: Size reduction. Academic Press, London, San Diego.

Bettencourt da Silva, R., A. Williams (Eds), 2015. Eurachem/CITAC Guide: Setting and Using Target Uncertainty in Chemical Measurement, 1st ed. 2015. Available from https://www.eurachem.org/index.php/publications/guides/gd-stmu

Bovera, F., M. Spanghero, G. Galassi, F. Masoero, A. Buccioni, 2003. Repeatability and reproducibility of the Cornell Net carbohydrate and Protein System analytical determinations. Ital. J. Anim. Sci. 2: 41-50.

Bremer, M.G.E.G., R.J.C.F. Margry, J.C.H. Vaessen, A.M.H. Van Doremalen, J.G.P. Van Der Palen, R.G.C. Van Kaathoven, A.E.M. Kemmers-Vonken, L.W.D. van Raamsdonk. 2013. Evaluation of a commercial ELISA for detection of ruminant processed animal proteins in non-ruminant processed animal proteins. J AOAC. 96:552-559.

Chow, Y. W., Pietranico, R., Mukerji, A., 1975. Studies of oxygen binding energy to hemoglobin molecule. Bioch. Bioph. Res. Comm 66: 1424–1431.

Ciasca B., M. Pascale, V.G. Altieri, F. Longobardi, M. Suman, D. Catellani, V.M.T. Lattanzio, 2018. In-house validation and small-scale collaborative study to evaluate analytical performances of multimycotoxin screening methods based on liquid chromatography–high-resolution mass spectrometry: Case study on Fusarium toxins in wheat. J. Mass Spectrom. 53: 743–752.

Codex Alimentarius, 2004. Guidelines on measurement uncertainty. CAC/GL 54-2004.

Codex Alimentarius, 2008. Guidelines for the validation of food safety control measures. CAC/GL 69 – 2008.

Currie, L.A., 1995. Nomenclature in evaluation of analytical methods including detection and quantification capabilities (IUPAC Recommendations 1995). Pure &Appl. Chem., Vol. 67, No. 10, pp. 1699-1723.

DG-SANCO, 2009. Guideline for initial validation and transfer of screening methods. Non-Paper 08/07/2009.

Ellison, S. L. R. and A. Williams, (Eds), 2012. Eurachem/CITAC guide number 4: Quantifying Uncertainty in Analytical Measurement, Third edition. ISBN 978-0-948926-30-3. Available from https://www.eurachem.org/index.php/publications/guides/quam

Engling, F.P., Jørgenson, J.S., Paradies-Severin, I., Hahn, H., 2000. Evidence of animal meal in feeds. FeedMagazine/Kraftfutter. 1: 14-17.

EPA, 2016. Method Validation of U.S. Environmental Protection Agency (EPA) Microbiological Methods of Analysis, prepared by The FEM Microbiology Action Team. FEM Document Number 2009-01, REVISION: December 21, 2016.

European Commission, 2002. Commission Decision (EC) 2002/657 of 14 August 2002 implementing Council Directive 96/23/EC concerning the performance of analytical methods and the interpretation of results. Off. J. Eur. Union, L 221, p. 8.

European Commission, 2013. Commission Regulation (EU) No 68/2013 of 16 January 2013 on the Catalogue of feed materials. Off. J. Eur. Union, L 29, 30.1.2013, p. 1–64.

European Commission, 2006. Regulation (EC) 401/2006 of 23 February 2006 laying down the methods of sampling and analysis for the official control of the levels of mycotoxins in foodstuffs. Off. J. Eur. Union, L 70, 9.3.2006, p. 12-53.

European Commission, 2017. Commission Regulation (EU) 2017/1017 of 15 June 2017 amending Regulation (EU) No 68/2013 on the Catalogue of feed materials. Off. J. Eur. Union, L 159, 21.6.2017, p. 48–119.
European Union, 2021. Commission Implementing Regulation (EU) 2021/808 of 22 March 2021 on the performance of analytical methods for residues of pharmacologically active substances used in food-producing animals and on the interpretation of results as well as on the methods to be used for sampling and repealing decisions 2002/657/EC and 98/179/EC. Off. J. Eur. Union, L 180, 21.5.2021, p. 84-114.
Fagerland, M.W., S. Lydersen, P. Laake, 2017. Contingency Tables. CRC Press, Taylor & Francis Group, Boca Raton.
FAO and WHO. 2019. Hazards associated with animal feed. Report of the Joint FAO/WHO expert meeting – 12–15 May 2015, FAO headquarters, Rome, Italy. FAO Animal Production and Health Report No. 13. Rome.
Fay, M.P. and M.A. Proschan, 2010. Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. Statistics Surveys 4: 1-39.
Feldsine, P.C. Abeyta, W.H. Andrews, 2002. Validation of Qualitative and Quantitative Food Microbiological Official Methods of Analysis. J AOAC, 85: 1187-1200.
Gizzi, G., Holst, C. von, Baeten, V., Berben, G., Raamsdonk, L.W.D. van, 2004. Determination of Processed Animal Proteins, Including Meat and Bone Meal, in Animal Feed. J. AOAC 87 (6): 1334-1341.
Godoy, M.R.C. de, M. Hervera, K.S. Swanson, G.C. Fahey Jr., 2016. Innovations in Canine and Feline Nutrition: Technologies for Food and Nutrition Assessment. Annu. Rev. Anim. Biosci. 4: 311–333.
González Medina, S., Hyde, C., Lovera, I., Piercy, R.J., 2018. Detection of equine atypical myopathy-associated hypoglycin A in plant material: Optimisation and validation of a novel LC-MS based method without derivatisation. PLoS ONE 13(7): e0199521.
Gower, J.C., 1971. A general coefficient of similarity and some of its properties. Biometrics, 27: 857-871.
Henneberg W, Stohmann F. 1859. Ueber das erhaltungsfutter volljaehrigen rindviehs. J. Landwirtsch. 3: 485–551.
Horwitz, W., 1995. Protocol for the design, conduct and interpretation of method-performance studies (IUPAC Technical Report). Pure & Appl. Chem. 67: 331-343.
Horwitz, W., R. Albert, 2006. The Horwitz Ratio (HorRat): A useful index of method performance with respect to precision. J AOAC International 89: 1095-1109.
Hubert, L., P. Arabie, 1985. Comparing partitions. J. Classification 2: 193-218.
Indira, T.N., S. Bhattacharya, 2006. Grinding characteristics of some legumes. Journal of Food Engineering 76 (2006) 113–118.
ISO, 1994a. Accuracy (trueness and precision) of measurement methods and results. Part 1. General principles and definitions. ISO 5725-1:1994(E). ISO, Geneva.
ISO, 1994b. Accuracy (trueness and precision) of measurement methods and results. Part 2. Basic method for the determination of repeatability and reproducibility of a standard measurement method. ISO 5725-2:1994(E). ISO, Geneva.
ISO, 1997. Proficiency Testing by Interlaboratory Comparisons, part 1: Development and operation of proficiency testing schemes, ISO/IEC 43-1:1997(E). ISO, Geneva.
ISO, 2003. Microbiology of food and animal feeding stuffs – Protocol for the validation of alternative method. ISO/FDIS 16140:2003(E). ISO, Geneva.
ISO, 2005. Statistical methods for use in proficiency testing by interlaboratory comparisons. International Standard Organisation, Geneva: ISO 13528:2015(E). ISO, Geneva.
ISO, 2006. Foodstuffs -- Methods of analysis for the detection of genetically modified organisms and derived products -- General requirements and definitions. ISO 24276:2006(E). ISO, Geneva.
ISO, 2010. Conformity assessment - General requirements for proficiency testing. ISO/IEC 17043:2010(E). CEN/CENELEC, Brussels.
ISO, 2017. General requirements for the competence of testing and calibration laboratories. ISO/ICE 17025:2017(E). ISO, Geneva.
ISO, 2019. Molecular biomarker analysis -- Methods of analysis for the detection and identification of animal species in foods and food products (nucleic acid-based methods) -- General requirements and definitions. ISO 20813:2019(E). ISO, Geneva.
IUPAC, 2014. Compendium of Chemical Terminology. Gold Book. https://goldbook.iupac.org/
JCGM 100:2008, 2010. Evaluation of measurement data — Guide to the expression of uncertainty in measurement. GUM 1995 with minor corrections. Working Group 1 of the Joint Committee for Guides in Metrology, https://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf

Käppler, A., D. Fischer, S. Oberbeckmann, M. Labrenz, K.-J. Eichhorn, B. Voit, 2016. Analysis of environmental microplastics by vibrational microspectroscopy: FTIR, Raman or both? Anal. Bioanal. Chem. 408: 8377–8391.

Karakolis E.G., B. Nguyen, J. Bem You, C.M. Rochman, D. Sinton, 2019. Fluorescent Dyes for Visualizing Microplastic Particles and Fibers in Laboratory-Based Studies. Environ. Sci. Technol. Lett. 6: 334-340.

Kiani, S., S.M. van Ruth, L.W.D. van Raamsdonk, S. Minaei, 2019. Hyperspectral imaging as a novel system for the authentication of spices: A nutmeg case study. Food Science and Technology 104: 61-69.

Koelmans, A.A., Mohamed Nor, N.H., Hermsen, E., Kooi, M., Mintenig, S.M., De France, J., 2019. Microplastics in freshwaters and drinking water: Critical review and assessment of data quality. Water Res 155: 410-422.

Korol, W., G. Bielecka, J. Rubaj, S. Walczyński, 2015. Uncertainty from sample preparation in the laboratory on the example of various feeds. Accred. Qual. Assur. 20: 61–66.

Langton, S.D., R. Chevennement, N. Nagelkerke & B. Lombard, 2002. Analysing collaborative trials for qualitative microbiological methods: accordance and concordance. International Journal of Food Microbiology, 79: 175-181.

Lee, Y.L., J.S. Yoo, W.B. Yoon, 2014. Grinding Characteristics of Black Soybeans (Glycine max) at Varied Moisture Contents: Particle Size, Energy Consumption, and Grinding Kinetics. International Journal of Food Engineering 10: 347–356

Liu X, Han LJ, Veys P, Baeten V, Jiang XP, Dardenne P. 2011. An Overview of the Legislation and Light Microscopy for Detection of Processed Animal Proteins in Feeds. Micr Res Technique. 74: 735-743.

Macarthur, R. and C. von Holst, 2012. A protocol for the validation of qualitative methods of detection. Anal. Methods 4: 2744-2754.

Maes, Th., R. Jessop, N. Wellner, K. Haupt, A.G. Mayes, 2017. A rapid-screening approach to detect and quantify microplastics based on fluorescent tagging with Nile Red. Scientific Reports 7: 44501. DOI: 10.1038/srep44501.

Marchis et al., 2021. Gravimetric quantitative validation of botanic impurities in feed. Journal of the Science of Food and Agriculture 101: 1047-1052.

McHugh, M.L., 2012. Interrater reliability: the kappa statistic. Biochemia Medica 22(3): 276-282.

Mulder, P.P.J., L.W.D. van Raamsdonk, H. van Egmond, T. van der Horst and J. de Jong, 2012. Ergot alkaloids in animal feed. Results of a survey in The Netherlands. Report 2012.005, RIKILT, Wageningen, pp. 50.

Naimi, L.J., F. Collard, X. Bi, C.J. Lim, S. Sokhansanj, 2016. Development of size reduction equations for calculating power input for grinding pine wood chips using hammer mill. Biomass Conv. Bioref. 6: 397–405.

Ohe, W. von der, L. Persano Oddo, M.L. Piana, M. Morlot, P. Martin, 2004. Harmonized methods of melissopalynology. Apidologie 35: S18–S25.

Ottoboni, M., M. Tretola, F. Cheli, D. Marchis, P. Veys, V. Baeten, L. Pinotti, 2017. Light microscopy with differential staining techniques for the characterisation and discrimination of insects versus marine arthropods processed animal proteins. Food Additives & Contaminants: Part A, 34: 1377-1383.

Pagter, E., Frias, J., Nash, R., 2018. Microplastics in Galway Bay: A comparison of sampling and separation methods. Mar. Pollut. Bull. 135: 932–940.

Pocklington, W.D., 1990. Harmonized protocols for the adoption of standardized analytical methods and for the presentation of their performance characteristics. Pure & Appl. Chern., Vol. 62, No. 1, pp. 149-162.

Raamsdonk, L.W.D. van, C.P.A.F. Smits, J. Vliege, H. van Egmond, 2022b. A validated method for detection and quantification of packaging material in candy syrups. Report 2022.008. WFSR, Wageningen, pp. 36.

Raamsdonk, L.W.D. van, G. Frick, I. Ujčič Vrhovnik, M. Zadravec, J. Zegers, R. Krull-Wöhrmann, R. Weiss, G. van der Borg, 2022a. Introduction to new guidelines for validation of methods to examine visually recognisable substances. Food Additives & Contaminants, Part A, https://doi.org/10.1080/19440049.2022.2135768

Raamsdonk, L.W.D. van, H. van der Voet, 2022b. Measurement uncertainty for detection of visual impurities in granular feed and food materials in relation to the investigated amount of material. Food Additives & Contaminants, Part A, 39: 1265-1283. https://doi.org/10.1080/19440049.2022.2066193

Raamsdonk, L.W.D. van, I.M.J. Scholtens, J. Ossenkoppele, H. van Egmond, M. Groot, 2011. Investigation into blood plasma in milk formula. Report 2011.003, RIKILT, Wageningen, pp. 19.

Raamsdonk, L.W.D. van, J.S. Jørgensen, P. Veys, J. Vancutsem, G. Pridotkas, 2012a. Classical microscopy - Improvements of the qualitative protocol. In: Detection, identification and quantification of processed animal proteins in feedingstuffs, Namur, Les éditions namuroises, Ch 5, 47-57. ISBN 978-2-87551-029-7.

Raamsdonk, L.W.D. van, L. Pinotti, P. Veys, A. Campagnoli, C. Paltanin, C. Belinchón Crespo, J.S. Jørgensen, 2012b. Markers for microscopic detection In: Detection, identification and quantification of processed animal proteins in feedingstuffs, Namur, Les éditions namuroises, Ch 6, 59-69. ISBN 978-2-87551-029-7.

Raamsdonk, L.W.D. van, M. van der Zande, A.A. Koelmans, L.A.P. Hoogenboom, R.J.B. Peters, M.J. Groot, A.A.C.M. Peijnenburg, Y.J.A Weesepoel, 2020. Current insights into monitoring, bioaccumulation, and potential health effects of microplastics present in the food chain. Foods 9: 72-99; doi:10.3390/foods9010072.

Raamsdonk, L.W.D. van, P. Mulder, M. Uiterwijk, 2010. Identification tools as part of Feedsafety research: the case of ragwort. In: P.L. Nimis and R. Vignes Lebbe, Proceedings of Bioidentify.eu: "Tools for identifying biodiversity: progress and problems", Edizioni Università di Trieste, pp. 213-216. http://www.openstarts.units.it/dspace/bitstream/10077/3776/1/van%20Raamsdonk,%20Mulder,%20Uiterwijk,%20bioidentify.pdf

Raamsdonk, L.W.D. van, R.J.C.F. Margry, R.G.C. van Kaathoven, M.G.E.G. Bremer, 2015. Inter-laboratory study of two immunochemical methods for detection of ruminant animal proteins. Food Chemistry, Volume 185, 15 October 2015: 333–339.

Raamsdonk, L.W.D. van, S. van der Vange, M. Uiterwijk, M.J. Groot, 2012d. Reliability and evaluation of identification models exemplified by a histological diagnosis model. In: Dr. Chiang Jao (ed.), Decision Support Systems, Intech, ISBN 980-953-307-529-2. http://dx.doi.org/10.5772/51362

Raamsdonk, L.W.D. van, T.W. Prins, N. van de Rhee, J.J.M. Vliege, V.G.Z. Pinckaers, 2017. Microscopic recognition and identification of fish meal in compound feeds. Food Additives and Contaminants Part A. 34: 1364-1376.

Raamsdonk, L.W.D. van, V. Pinckaers, J. Vliege, H. van Egmond, 2012c. Examination of packaging materials in bakery products. A validated method for detection and quantification. Report 2012.007, RIKILT, Wageningen, pp. 20.

Rand, W.M., 1971. Objective Criteria for the Evaluation of Clustering Methods. J. Am. Stat. Assoc. 66: 846-850.

Reiter, E., Zentek, J, Razzazi, E., 2009. Review on sample preparation strategies and methods used for the analysis of aflatoxins in food and feed. Mol. Nutr. Food Res. 53: 508 – 524.

Reiter, E.V., Dutton, M.F., Agus, A., Nordkvist, E., Mwanza, M.F., Njobeh, P.B., Prawano, D., Häggblom, P., Razzazi-Fazeli, E., Zentek, J., Andersson, M.G., 2011. Uncertainty from sampling in measurements of aflatoxins in animal feedingstuffs: application of the Eurachem/CITAC guidelines. Analyst, 136: 4059-4069.

Rijk, T.C. de, R.C.J. van Dam, P. Zomer, E.A.M. Boers, P. de Waard, H.G.J. Mol, 2013. Development and validation of a confirmative LC-MS/MS method for the determination of ß-exotoxin thuringiensin in plant protection products and selected greenhouse crops. Anal. Bioanal. Chem. 405:1631–1639.

Sedgwick, P., 2014. Pitfalls of statistical hypothesis testing: type I and type II errors. BMJ 349: g4287. doi: 10.1136/bmj.g4287

Shashidhar, M.G., T.P. Krishna Murthy, K.G. Girish, B. Manohar, 2013. Grinding of Coriander Seeds: Modeling of Particle Size Distribution and Energy Studies. Particulate Science and Technology, 31:5, 449-457,

Sheskin, D., 2004. Handbook of parametric and nonparametric statistical procedures. Chapman & Hal/CRC, Boca Raton, Florida.

Sneath, P.H.A., R.R. Sokal, 1973. Numerical taxonomy. The principles and practice of numerical classification. W.H. Freeman and Company, San Francisco.

Stehman, S.V., 1997. Selecting and interpreting measures of thematic classification accuracy. Remote Sens. Environm. 62: 77—89.

Thompson, M., R. Wood, 1995. Harmonized guidelines for internal quality control in analytical chemistry laboratories. Pure & Appl. Chem., Vol. 67, No. 4, pp. 649-666.

Thompson, M., S.L.R. Ellison, R. Wood, 2002. Harmonized guidelines for single laboratory validation of methods of analysis (IUPAC Technical Report). Pure Appl. Chem., Vol. 74, No. 5, pp. 835–855.

Uiterwijk, M., L.W.D. van Raamsdonk, S.J.C. Janssen, 2013. Determinator – A Generic DSS For Hazard Identification Of Species Or Other Physical Subjects. In: R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp (Eds.): International Environmental Modelling and Software Society (iEMSs) 2012 International Congress

on Environmental Modelling and Software Managing Resources of a Limited Planet: Pathways and Visions under Uncertainty, Sixth Biennial Meeting, Leipzig, Germany, p. 581-587. https://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=1740&context=iemssconference

Veys, P., Baeten, V. 2018. Protocol for the isolation of processed animal proteins from insects in feed and their identification by microscopy. Food Control. 92: 496–504.

Veys, P., Berben, G., Dardenne, P., Baeten, V., 2012. Detection and identification of animal by-products in animal feed for the control of Transmissible Spongiform Encephalopathies. In: Animal Feed Contamination. Cambridge, United Kingdom: Woodhead Publishing Ltd.

Veys, P., V. Planchon, R. Colbert, C. Cruz, G. Frick, I. Ioannou, D. Marchis, E. Nordkvist, I. Paradies-Severin, A. Pohto, R. Weiss, V. Baeten & G. Berben, 2017. Collaborative study on the effect of grinding on the detection of bones from processed animal proteins in feed by light microscopy. Food Additives & Contaminants: Part A, 34: 1451-1460.

Voet, H. van der, and L.W.D. van Raamsdonk, 2004. Estimation of accordance and concordance in inter-laboratory trials of analytical methods with qualitative results. Intl. J. of Food Microbiol. 95: 231-234.

Wehling, P., R.A. Labudde, S. L. Brunelle, M. T. Nelson, 2011. Probability of Detection (POD) as a Statistical Model for the Validation of Qualitative Methods. J. AOAC Int. 94 (1): 335-347.

Wheeler, Donald J., 2000. Understanding Variation. Knoxville, Tennessee: SPC Press. ISBN 978-0-945320-53-1.

Handbooks for visual identification

Alonso, G. L., M. R. Salinas, J. Garuo, 1998. Method to Determine the Authenticity of Aroma of Saffron (*Crocus sativus* L.). Journal of Food Protection, Vol. 61, No. 11, 1998, Pages 1525-1528.

Cappers, R.T.J., R.M. Bekker, J.E.A. Jans, 2012. Digitale Zadenatlas van Nederland / Digital Seed Atlas of the Netherlands. Groningen Archaeological Studies 4, Barkhuis, Groningen. https://www.plantatlas.eu/

Czaja, A., 1969. Die Mikroskopie der Stärkekörner. Paul Parey, Berlin.

Eschrig, W., 1999. Pulver-Atlas de Drogen der deutschsprachigen Arzneibücher. Deutscher Apotheker Verlag, Stuttgart.

Feigl, F., V. Anger, 1958. Spot Tests in Inorganic Analysis, 6th Edition. Elsevier.

Fischer, R. en Th. Kartnig, 1978. Drogenanalyse. Makroskopische und mikroskopische Drogenuntersuchungen. Peinger Verlag, Wien, New York.

Flint, O., 1994. Food microscopy: a manual of practical methods, using optical microscopy. Bios Scientific Publishers / Royal Microscopical Society.

Gassner, G., 1973. Mikroskopische Untersuchungen pflanzlicher Lebensmittel. Gustav Fischer Verlag, Stuttgart.

Gassner, G., B. Hohmann, F. Deutschmann, 1989. Mikroskopische Untersuchungen pflanzlicher Lebensmittel, 5. Auflage. Gustav Fischer Verlag, Stuttgart.

Hahn, H., I. Michaelsen, 1996. Mikroskopische Diagnostik pflanzlicher Nahrungs-, Genuss- und Futtermittel, einschliesslich Gewuerze. Springer Verlag, Berlin.

Hohmann, B., 2006. Mikroskopische Untersuchung Pflanzlicher Lebensmittel Und Futtermittel. Behr's Verlag, Hamburg.

Hohmann, B., G. Reher, E. Stahl-Biskup, 2001. Mikroskopische Drogenmonographien der deutschsprachigen Arzneibücher. Wissenschaftliche Verlagsgesellschaft mbH, Stuttgart.

ISO, 2010. Spices – Saffron (*Crocus sativus* L.) – Part 2: Test methods, NEN-ISO 3632-2:2010.

ISO, 2011. Spices – Saffron (*Crocus sativus* L.) – Part 1: Specification, NEN-ISO 3632-1:2011.

Mészáros, L., Bihler, E., 1983. Atlas für die Mikroskopie von Nahrungsgrundstoffen und Futtermitteln. Teil II: Stärkereiche Nahrungsgrundstoffe und deren Verarbeitungsprodukte, Grünmehle, Obstrester, Braunalgen u.a. Verlag J. Neumann-Neudamm, Melsungen.

Rahfeld, B., 2009. Mikroskopischer Farbatlas pflanzlicher Drogen. Spektrum, Heidelberg.

Sawyer, R., 1981. Pollen Identification for Beekeepers. Northern Bee Books. ISBN 978-1904846062.

Sawyer, R., 1988. Honey identification. Cardiff Academic Press. ISBN 1-871254-00-0.

Seidemann, J., 1966. Stärke-Atlas. Parey, Berlin.

Wichtl, M., 1997. Teedrogen und Phytopharmaka. Wissenschaftliche Verlagsgesellschaft mbH, Stuttgart.

Winton, A.L., Barber Winton, K., (1932) 1946. The structure and composition of foods. Volume I: cereals, starch, oil seeds, nuts, oils, forage plants. John Wiley & Sons, New York.

Winton, A.L., Barber Winton, K., 1939. The structure and composition of foods. Volume IV: sugar, cocoa, coffee, tea, spices, extracts, yeast, baking powder. John Wiley & Sons, New York.

Proficiency testing visual methods

IAG animal proteins

Raamsdonk, L.W.D. van, Hekman, W., Vliege, J.M., Pinckaers, V., Voet, H. van der, Ruth, S.M. van, 2008. The 2008 Dutch NRL / IAG proficiency test for detection of animal proteins in feed. Report 2008.007, RIKILT, Wageningen, 31 pp.

Raamsdonk, L.W.D. van, Hekman, W., Vliege, J.M., Pinckaers, V., Ruth, S.M. van, 2009. Animal proteins in feed. IAG ring test 2009. Report 2009.017, RIKILT, Wageningen, 34 pp.

Raamsdonk, L.W.D. van, Hekman, W., Vliege, J.M., Pinckaers, V., Ruth, S.M. van, 2010. Animal proteins in feed. IAG ring test 2010. Report 2010.009, RIKILT, Wageningen, pp. 38.

Raamsdonk, L.W.D. van, Pinckaers, V., Vliege, J.M., Ruth, S.M. van, 2011. Animal proteins in feed. IAG ring test 2011. Report 2011.015, RIKILT, Wageningen, pp. 38.

Raamsdonk, L.W.D. van, Pinckaers, V.G.Z., Vliege, J.J.M., 2012. Animal proteins in feed. IAG ring test 2012. Wageningen: WUR, report RIKILT 2012.009.

Raamsdonk, L.W.D. van, Pinckaers, V.G.Z., Vliege, J.J.M., 2013. Animal proteins in feed. IAG ring test 2013. Report 2013.016. Wageningen: RIKILT, p. 35.

Raamsdonk, L.W.D. van, Pinckaers, V.G.Z., Scholtens, I.M.J., Prins, T.W., Voet, H. van der, Vliege, J.J.M., 2014. IAG ring test animal proteins 2014. Report 2014.011. Wageningen: RIKILT, p. 47.

Raamsdonk, L.W.D. van, Rhee, N. van de, Scholtens, I.M.J., Prins, T.W., Vliege, J.J.M., Pinckaers, V.G.Z., 2015. IAG ring test animal proteins 2015. Report 2015.016. Wageningen: RIKILT, p. 32.

Raamsdonk, L.W.D. van, Rhee, N. van de, Scholtens, I.M.J., Prins, T.W., Vliege, J.J.M., Pinckaers, V.G.Z., 2016. IAG ring test animal proteins 2016. Report 2016.008. Wageningen: RIKILT, p. 31.

Raamsdonk, L.W.D. van, J.J.M. Vliege, C.P.A.F. Smits, V.G.Z. Pinckaers, 2017. IAG ring test animal proteins 2017. Report 2017.012. RIKILT, Wageningen, pp. 34.

Raamsdonk, L.W.D. van, B. Hedemann, C.P.A.F. Smits, J.J.M. Vliege, 2018. IAG ring test animal proteins 2018. Report 2018.008. RIKILT, Wageningen, pp. 33.

Raamsdonk L.W.D. van, C.P.A.F. Smits, B. Hedemann, T.W. Prins, J.J.M. Vliege, 2019. IAG proficiency test animal proteins 2019. Report 2019.015. WFSR, Wageningen, pp. 36.

Raamsdonk L.W.D. van, C.P.A.F. Smits, B. Hedemann, T.W. Prins, 2021. IAG proficiency test animal proteins 2021. Report 2021.019. WFSR, Wageningen, pp. 36.

Borg, G. van der, C.P.A.F. Smits, B. Hedemann, T.W. Prins, L.W.D. van Raamsdonk, 2022. IAG proficiency test animal proteins 2022. Report 2022.010. WFSR, Wageningen, pp. 35.

EURL animal proteins

Veys, P., V. Baeten, 2007. CRL-AP Proficiency Test 2006. Final Report. Community Reference Laboratory for Animal Proteins in feedingstuffs, Walloon Agricultural Research Centre, Gembloux, Belgium.

Veys, P., G. Berben, V. Baeten, 2007. CRL-AP Proficiency Test 2007. Final Report. Community Reference Laboratory for Animal Proteins in feedingstuffs, Walloon Agricultural Research Centre, Gembloux, Belgium.

Veys, P., V. Baeten, 2008. CRL-AP Interlaboratory Study 2007. Final Report. Community Reference Laboratory for Animal Proteins in feedingstuffs, Walloon Agricultural Research Centre, Gembloux, Belgium.

Veys, P., G. Berben, V. Baeten, 2009. CRL-AP Proficiency Test 2008. Final Report. Community Reference Laboratory for Animal Proteins in feedingstuffs, Walloon Agricultural Research Centre, Gembloux, Belgium.

Veys, P., G. Berben, V. Baeten, 2010. CRL-AP Proficiency Test 2009. Final Report. Community Reference Laboratory for Animal Proteins in feedingstuffs, Walloon Agricultural Research Centre, Gembloux, Belgium.

Veys, P., G. Berben, V. Baeten, 2011. EURL-AP Interlaboratory Study Microscopy 2010. Validation of a revised version of Annex VI of EU Regulation EC/152/2009 and proficiency evaluation. Final report. CRAW, Gembloux.

Veys, P., Berben, G., Baeten, V., 2011. EURL-AP Interlaboratory Study Microscopy 2010. Final version. European Union Reference Laboratory for Animal Proteins in feedingstuffs, Walloon Agricultural Research Centre, Gembloux, Belgium.
Veys, P., Berben, G., Baeten, V., 2012. EURL-AP Proficiency Test Microscopy 2011. Final version. European Union Reference Laboratory for Animal Proteins in feedingstuffs, Walloon Agricultural Research Centre, Gembloux, Belgium.
Veys, P., Berben, G., 2013. EURL-AP Proficiency Test Microscopy 2012. Final version. European Union Reference Laboratory for Animal Proteins in feedingstuffs, Walloon Agricultural Research Centre, Gembloux, Belgium.
Veys, P., Baeten, V., Berben, G., 2014. EURL-AP Proficiency Test Microscopy 2013. Final version. European Union Reference Laboratory for Animal Proteins in feedingstuffs, Walloon Agricultural Research Centre, Gembloux, Belgium.
Veys, P., Baeten, V., Berben, G., 2015. EURL-AP Proficiency Test Microscopy 2014. Final version. European Union Reference Laboratory for Animal Proteins in feedingstuffs, Walloon Agricultural Research Centre, Gembloux, Belgium.
Veys, P., Fumière, O., Marien, A., Baeten, V., Berben, G., 2016. Combined microscopy-PCR EURL-AP Proficiency Test 2015. Final version. European Union Reference Laboratory for Animal Proteins in feedingstuffs, Walloon Agricultural Research Centre, Gembloux, Belgium.
Fumière, O., Veys, P., Marien, A., Baeten, V., Berben, G., 2017. Combined microscopy-PCR EURL-AP Proficiency Test 2016. Final version. European Union Reference Laboratory for Animal Proteins in feedingstuffs, Walloon Agricultural Research Centre, Gembloux, Belgium.
Veys, P., Fumière, O., Marien, A., Baeten, V., Berben, G., 2018. Combined microscopy-PCR EURL-AP Proficiency Test 2017. Final version. European Union Reference Laboratory for Animal Proteins in feedingstuffs, Walloon Agricultural Research Centre, Gembloux, Belgium.
Fumière, O., Veys, P., Marien, A., Baeten, V., Berben, G., 2019. Combined microscopy-PCR EURL-AP Proficiency Test 2018. Final version. European Union Reference Laboratory for Animal Proteins in feedingstuffs, Walloon Agricultural Research Centre, Gembloux, Belgium.
Veys, P., Fumière, O., Marien, A., 2020. Combined microscopy-PCR EURL-AP Proficiency Test 2019. Final version. European Union Reference Laboratory for Animal Proteins in feedingstuffs, Walloon Agricultural Research Centre, Gembloux, Belgium.
Fumière, O., P. Veys & A. Marien, 2021. Combined microscopy-PCR EURL-AP Proficiency Test 2020. Final version. European Union Reference Laboratory for Animal Proteins in feedingstuffs, Walloon Agricultural Research Centre, Gembloux, Belgium.
Veys, P., Fumière, O., 2021. Implementation study on the detection of insect PAP in feed by double sedimentation method PE/TCE followed by light microscopy detection. European Union Reference Laboratory for Animal Proteins in feedingstuffs, Walloon Agricultural Research Centre, Gembloux, Belgium.
Veys, P., Fumière, O., Marien, A., 2022. Combined microscopy-PCR EURL-AP Proficiency Test 2021. Final version. European Union Reference Laboratory for Animal Proteins in feedingstuffs, Walloon Agricultural Research Centre, Gembloux, Belgium.

IAG composition

Raamsdonk, L.W.D. van, Pinckaers, V.G.Z., Vliege, J.J.M., 2014. IAG ring test feed composition 2014. Report 2014.010. RIKILT, Wageningen, pp. 23.
Raamsdonk, L.W.D. van, Rhee, N. van de, Pinckaers, V.G.Z., Vliege, J.J.M., 2015. IAG ring test feed composition 2015. Report 2015.017. RIKILT, Wageningen, pp. 26.
Raamsdonk, L.W.D. van, Rhee, N. van de, Pinckaers, V.G.Z., Vliege, J.J.M., 2016. IAG ring test feed composition 2016. Report 2016.014. RIKILT, Wageningen, pp. 26.
Raamsdonk, L.W.D. van, C.P.A.F. Smits, J.J.M. Vliege, V.G.Z. Pinckaers, 2017. IAG ring test feed composition 2017. Report 2017.011. RIKILT, Wageningen, pp. 32.
Raamsdonk L.W.D. van, C.P.A.F. Smits, J.J.M. Vliege, 2020. IAG proficiency test feed composition 2019. Report 2019.014. WFSR, Wageningen, pp. 26.
Raamsdonk, L.W.D. van, C.P.A.F. Smits, B. Hedemann, 2022. IAG proficiency test feed composition 2021. Report 2022.021. WFSR, Wageningen, pp. 27.

IAG other

Raamsdonk, L.W.D. van, N. van de Rhee, J.M. Vliege, V. Pinckaers, 2016. IAG ring test visual detection of ergot sclerotia in rye 2015. Report 2016.013. RIKILT, Wageningen, pp. 26.

Raamsdonk, L.W.D. van, C.P.A.F. Smits, B. Hedemann, 2020. Proficiency test of detection of packaging material in bakery by-products 2019. Report 2020.022. WFSR, Wageningen, pp. 20.

Peereboom, D.P.K.H., J.B.G.M. Hedemann, C.P.A.F. Smits, M. de Nijs, L.W.D. van Raamsdonk, 2021. Proficiency test for ergot sclerotia in cereals; EURLPT-MP05 (2020). Report 2021.002, WFSR, Wageningen, 34 pp.

Raamsdonk, L.W.D. van, C.P.A.F. Smits, B. Hedemann, G. van der Borg, 2022. Proficiency test of detection of packaging material in bakery by-products 2021. Report 2022.009, WFSR, Wageningen, 29 pp.

Other

CLEN, 2018. Proficiency test on Rice. Final Report, revised version CLEN Action 2. TAXUD, Brussels, contract TAXUD/2017/DE/306.

Gizzi G, von Holst C, Baeten V, Berben G, van Raamsdonk LWD. 2004. Determination of Processed Animal Proteins, Including Meat and Bone Meal, in Animal Feed. J AOAC. 87(6): 1334-1341.

Raamsdonk, L.W.D. van, H. van der Voet, 2003. A ring trial for the detection of animal tissues in feeds in the presence of fish meal. Report 2003.012, RIKILT, Wageningen. pp. 17 with 6 addenda.

Websites

IAG website: https://www.iag-micro.org/

All website links have been checked on 27 September 2022.

Appendix 1 Confidence intervals

Quantitative methods

A major parameter to express precision is measurement uncertainty (ISO/IEC 17025:2017 Chapter 7.6 (Evaluation of measurement uncertainty). This parameter is usually based on several factors: intralaboratory reproducibility, inhomogeneity and the bias from the true value are the most important elements (Ellison and Williams, 2012 (Eurachem QUAM:2012)). A relationship has been established between the level of contamination and the accepted maximum values (Pocklington, 1990; AOAC, 2002; Codex Alimentarius, 2004; Horwitz and Albert, 2006). Regulation (EU) 2021/808 presents some values adapted from the Horwitz equation in Table 2. The upper confidence limit as defined by Pocklington and AOC is based on the equation $RSD_R = 2 \cdot C^{-0.15}$, whereas Codex Alimentarius used a factor 4 ($RSD_R = 4 \cdot C^{-0.15}$). Chemical studies show levels of measurement uncertainty which are lower than the accepted maximum values according to the standard models (De Rijk et al., 2013; Ciasca et al., 2018; González Medina et al., 2018). The models and values from practice are shown in Figure A1.

Inhomogeneity has a small share in measurement uncertainty for chemical methods. In order to find the levels to be expected for the detection of visual units, WFSR has investigated samples of bird feed and whole grain cereals from practice for their level of contamination with Ambrosia seeds and Ergot sclerotia during the period of 2016 until 2019. Weight uncertainty has been calculated as the relative standard deviation under divisibility circumstances (see Chapter 4.1.5), since bias cannot be calculated for the lack of a true value (samples from practice). Samples were split in four equally sized subsamples using a splitter for achieving maximum homogeneity. The procedure is followed as set out in Chapter Quality control of sample analyses (Figure 3). All four subsamples were analysed during the period of this trial. Data from CEN 15587:2018 and from Denmark (unpublished) have been added. The upper confidence limit of the distribution of the values for RSD_d equals the equation $RSD_d = 2 \cdot C^{-0.41}$. This is a first attempt to develop documentation for a dedicated approach of measurement uncertainty for quantitative visual methods.

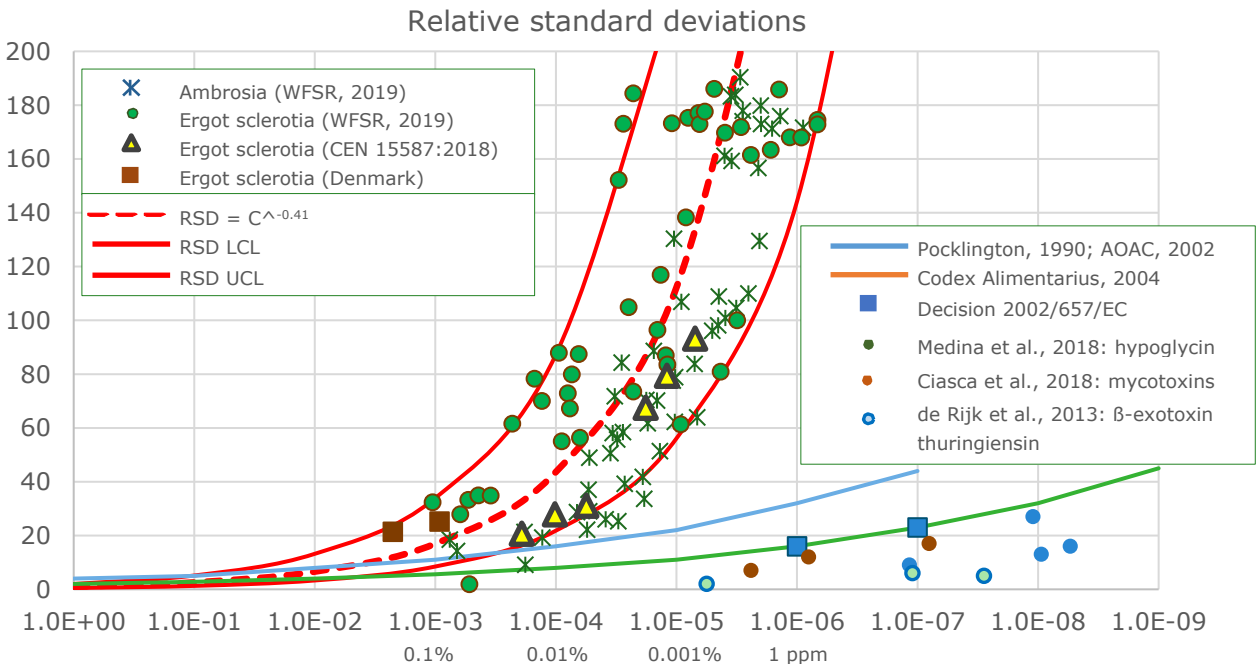


Figure A1 *Inhomogeneity among subsamples expressed as relative standard deviations.*

Qualitative methods

Two different approaches for establishing the probability of detection for qualitative results will be discussed. Macarthur and von Holst (2012) apply an inverse Beta distribution for calculating the confidence limits, whereas Wehling et al. (2011) apply a normal approximation of the binomial probabilities.

Let us assume a design with 12 participants which analysed 10 samples each (N_{lab} = 12 and $m_{samples}$ = 10). Two different scenarios have been evaluated with an increasing number of incorrect results: a) the incorrect results are evenly distributed over the participants, and b) the incorrect results are all reported by one participant. The table below shows the two scenarios (top half: a), and bottom half: b)). Parameter SE is calculated as $TP / (TP + FN)$.

total	correct	*SE*	1	2	3	4	5	6	7	8	9	10	11	12
120	120	1	1	1	1	1	1	1	1	1	1	1	1	1
120	119	0.9917	0.9	1	1	1	1	1	1	1	1	1	1	1
120	118	0.9833	0.9	0.9	1	1	1	1	1	1	1	1	1	1
120	117	0.975	0.9	0.9	0.9	1	1	1	1	1	1	1	1	1
120	116	0.9667	0.9	0.9	0.9	0.9	1	1	1	1	1	1	1	1
120	115	0.9583	0.9	0.9	0.9	0.9	0.9	1	1	1	1	1	1	1
120	120	1	1	1	1	1	1	1	1	1	1	1	1	1
120	119	0.9917	0.9	1	1	1	1	1	1	1	1	1	1	1
120	118	0.9833	0.8	1	1	1	1	1	1	1	1	1	1	1
120	117	0.975	0.7	1	1	1	1	1	1	1	1	1	1	1
120	116	0.9667	0.6	1	1	1	1	1	1	1	1	1	1	1
120	115	0.9583	0.5	1	1	1	1	1	1	1	1	1	1	1

The identical scenarios in terms of Accuracy turn out to provide identical upper limits for both approaches. The lower limit, however, is different. Macarthur and von Holst (2012) calculate lower and upper limits of the observed between-laboratory variation (L_s and U_s) as well as from the pure sampling error (L_H and U_H). The lower limit of the 95% confidence interval is equal to minimum (L_s | L_H), and the upper limit is equal to maximum (U_s | U_H). The values based on the observed between-laboratory variation turn out to give a larger range, as is illustrated in Figure A2 and A3. In scenario b) (all incorrect results are produced by one participant) the lower limit according to Macarthur and von Holst (2012) is directly related to that number of incorrect results. Based on this evaluation the participant with repetitive incorrect results could be removed from the dataset as outlying (underperforming). The upper confidence limits of both approaches for both scenarios turn out to be equal. The value of U_s (the upper confidence limit) remains at the maximum value of 1, independent of the number of incorrect results. The approach of Macarthur and von Holst (2012) was applied to the detection of animal proteins in feed. Both approaches were applied in an ILS of two immunoassays for the detection of ruminant troponin, which both provide an intermediate quantitative result, translated to a final qualitative result, which is, in respect to the data structure, comparable to the microscopic method for the detection of animal proteins in feed (van Raamsdonk et al., 2015).

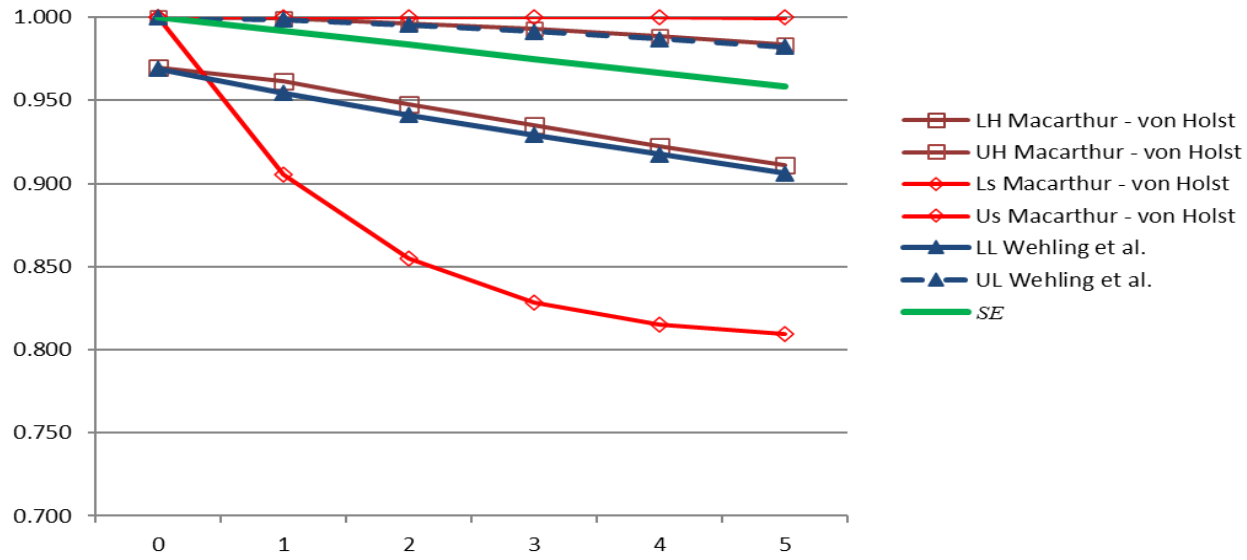


Figure A2 *Values for SE and upper and lower limits according to three approaches for zero, one, two etc. errors each made by different participants in a PT. For further explanation see text.*

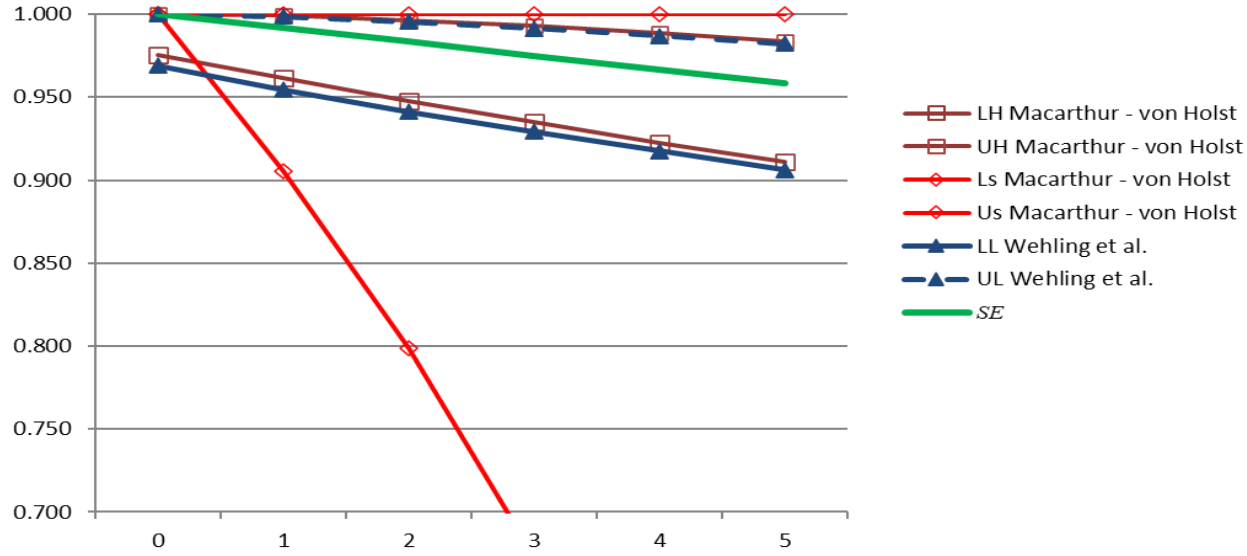


Figure A3 *Values for SE and upper and lower limits according to three approaches for zero, one, two etc. errors each made by the same participant in a PT. For further explanation see text.*

Equations applied:

Macarthur and von Holst (2012)

$$L_s = InverseBeta\left(0.025, \bar{p} \cdot \left(\left(\frac{\bar{p}\cdot(1-\bar{p})}{s^2}\right) - 1\right), \bar{p} \cdot \frac{(1-\bar{p})}{\bar{p}} \cdot \left(\left(\frac{\bar{p}\cdot(1-\bar{p})}{s^2}\right) - 1\right)\right) \quad \text{(A1)}$$

$$U_s = InverseBeta\left(0.975, \bar{p} \cdot \left(\left(\frac{\bar{p}\cdot(1-\bar{p})}{s^2}\right) - 1\right), \bar{p} \cdot \frac{(1-\bar{p})}{\bar{p}} \cdot \left(\left(\frac{\bar{p}\cdot(1-\bar{p})}{s^2}\right) - 1\right)\right) \quad \text{(A2)}$$

$$L_H = InverseBeta(0.025, (X + 0.5), (N - X + 0.5)) \quad \text{(A3)}$$

$$L_H = InverseBeta(0.975, (X + 0.5), (N - X + 0.5)) \quad \text{(A4)}$$

Wehling et al. (2011)

$$POD = x/N \tag{A5}$$

$$LCL = \frac{x+1.9207-1.9600\cdot\sqrt{x-\frac{x^2}{N}+0.9604}}{N+3.8415} \tag{A6}$$

$$UCL = \frac{x+1.9207+1.9600\cdot\sqrt{x-\frac{x^2}{N}+0.9604}}{N+3.8415} \tag{A7}$$

Appendix 2 Definitions of terms

Definitions are taken from Regulation (EU) 2021/808, Guidelines or paters as far as reasonable for the purpose of this Guidance. References are included in brackets; comments are added in square brackets when necessary; a number in brackets at the start of a definition refers to the definition in Article 2 of Regulation (EU) 2021/808. Lack of reference indicates definitions developed for the current Guidance (in **Bold**).

Accordance	The (percentage) chance that two identical test materials analysed by the same laboratory under standard repeatability conditions will both be given the same result (i.e. both found positive or both found negative). This is equivalent to repeatability for quantitative results. (Langton et al., 2002)
Absolute recovery	(1) the yield of the final stage of an analytical process for an analyte divided by the amount of the analyte in the original sample, expressed as a percentage. (Regulation (EU) 2021/808)
Accuracy	(2) the closeness of agreement between a test result and the accepted true reference value, determined by estimating trueness and precision. (Regulation (EU) 2021/808; ISO 24276:2006) An indication of a class of quality parameters. This class includes trueness/recovery, sensitivity (detection limit) and specificity, among some other additional parameters. The former name of the combined parameter for sensitivity and specificity for qualitative methods. Replaced by **Correctness**.
Alpha (*α*) error	(3) the probability that the tested sample is compliant, even though a non-compliant measurement result has been obtained. (Regulation (EU) 2021/808)
Analyte	(4) the component of a system to be analysed. (Regulation (EU) 2021/808)
Analytical threshold	A level of contamination above which a second (set of) analysis has to be carried out.
Applicability	Scope of application of the method which identifies the matrix, analyte or species being measured, its concentration range and the type of study/monitoring effort for which the procedure, as judged from its performance characteristics, is suited. Synonym to fitness for purpose. (ISO 24276:2006)
Authorised substance	(5) a pharmacologically active substance authorised for use in food-producing animals in accordance with Directive 2001/82/EC of the European Parliament and of the Council. (Regulation (EU) 2021/808)
Beta (*β*) error	(6) the probability that the tested sample is truly non-compliant, even though a compliant measurement result has been obtained. (Regulation (EU) 2021/808)
Bias	(7) the difference between the estimated value of the test result and an accepted reference value. (Regulation (EU) 2021/808) [This can be used as alternative term for trueness] Difference between the expectation of the test results and an accepted reference value. Bias is the total systematic error as contrasted to random error. There may be one or more systematic error components contributing to the bias. (AOAC, 2016)
Calibration standard	(8) a traceable reference for measurements that represents the quantity of substance of interest in a way that ties its value to a reference base. (Regulation (EU) 2021/808)
Certified reference material (CRM)	(9) a reference material, accompanied by documentation issued by a delegated body and providing one or more specified property values with associated uncertainties and traceabilities, using valid procedures. (Regulation (EU) 2021/808)
Collaborative study	(11) analysing the same sample(s) by using the same method to determine performance characteristics of the method in different laboratories, where the study allows to calculate the random measurement error and laboratory bias for the method used. (Regulation (EU) 2021/808)
Concordance	The percentage chance that two identical test materials sent to different laboratories will both be given the same result (i.e. both found positive or both found negative result). This is equivalent to reproducibility for quantitative results. (Langton et al., 2002)
Confirmatory method	(12) a method that provides full or complementary information enabling the substance to be unequivocally identified and if necessary quantified in one of the following manners: (a) at the maximum residue level or maximum level for authorised substances; (b) at the reference points for action (RPA) for prohibited or unauthorised substances, for which a reference point for action is established; (c) at a concentration as low as reasonably achievable for prohibited or unauthorised substance, for which no reference point for action is established. (Regulation (EU) 2021/808)

Correctness	Indication of the frequency of correct results obtained by examination of a set of identical samples. This parameter pertains to qualitative results and is expressed as percentage. It is calculated from the frequency of correct positive and correct negative results over the total number of results as denominator. The alternative term Accuracy (AC) has been used in a range of interlaboratory studies of visual monitoring methods, but is replaced by Correctness to avoid confusion.
Count	The number of units as principal result of a visual monitoring method. In quantitative methods the units will be weighted to deliver the derived result as percentage (w/w) for enforcement of the permitted limit.
Count dispersal	The diversity among counts as resulting from investigating replicate test portions.
Decision limit for confirmation (CC*α)*	(14) the limit at and above which it can be concluded with an error probability of α that a sample is non-compliant and the value 1 – α means statistical certainty in percentage that the permitted limit has been exceeded. (Regulation (EU) 2021/808)
Detection capability for screening (CC*β)*	(15) the smallest content of the analyte that may be detected or quantified in a sample with an error probability of β: (a) in the case of prohibited or unauthorised pharmacologically active substances, the CCβ is the lowest concentration at which a method is able to detect or quantify, with a statistical certainty of 1 – β, samples containing residues of prohibited or unauthorised substances; (b) in the case of authorised substances, the CCβ is the concentration at which the method is able to detect concentrations below the permitted limit with a statistical certainty of 1 – β. (Regulation (EU) 2021/808)
Detectability	The reliability of a positive identification of a particle of a contaminant or legal ingredient, expressed as a percentage. See Expert System.
Environment control	Control used to determine that there is no nucleic acid contamination from, for example, the air in the laboratory. (ISO 24276:2006)
Exclusivity	Exclusivity or specificity is the lack of interference in the alternative method from a relevant range of nontarget strains, which are potentially cross-reactive. (Feldsine et al., 2002 (AOAC)) Strains or isolates or variants of the target agent(s) that the method must not detect. (AOAC, 2016)
Expert system	A computer-based interactive system providing support to the process of identification. Characteristics of the particle found can be entered, resulting in match percentages with the targets in the library of the system (free access key). Comparable to libraries with chemical or DNA data for identification of results from MS or PCR.
False negative	An incorrect indication of the absence of a contaminant or undesirable substance, or an incorrect report of a number of particles below a chosen threshold. This definition is applicable to both quantitative and qualitative examination methods. An incorrect result below the legal limit; incorrect compliance.
False positive	An incorrect indication of the presence of a contaminant or undesirable substance, or an incorrect report of a number of particles exceeding a chosen threshold. This definition is applicable to both quantitative and qualitative examination methods. An incorrect result exceeding the legal limit; incorrect non-compliance.
Fortified sample material	(16) a sample enriched with a known amount of the analyte to be detected or quantified. (Regulation (EU) 2021/808)
Inclusivity	Inclusivity or sensitivity is the ability of the alternative method to detect the target analyte from a wide range of strains. (Feldsine et al., 2002 (AOAC)) Strains or isolates or variants of the target agent(s) that the method can detect. (AOAC, 2016)
Interlaboratory study (comparison)	(17) The organisation, performance and evaluation of tests on the same sample(s) by two or more laboratories in accordance with predetermined conditions to evaluate testing performance, either as a collaborative study or a proficiency test. (Regulation (EU) 2021/808) Study in which several laboratories detect and/or determine an analyte in one or more "identical" portions of homogeneous, stable materials under documented conditions. (ISO 24276:2006)
Internal standard	(18) a substance not contained in the sample and having physico-chemical properties as similar as possible to those of the analyte to be identified or quantified. (Regulation (EU) 2021/808)
Matrix	(21) the material from which a sample is taken. (Regulation (EU) 2021/808)
Matrix effect	(22) the difference in analytical response between a standard dissolved in the solvent and a matrix-matched standard either without a correction using an internal standard or with correction using an internal standard. (Regulation (EU) 2021/808)
measurand	(25) the particular quantity subject to measurement. (Regulation (EU) 2021/808) [Measurand is an alternative indication of contaminant or target]

Measurement uncertainty	(26) a non-negative parameter associated with the result of measurement, which characterises the dispersion of values that could reasonably be attributed to the measurand, based on the information used. (Regulation (EU) 2021/808) Parameter, associated with the result of a measurement, that characterises the dispersion of the values that could reasonably be attributed to the measurand. (Guide to the Expression of Uncertainty in Measurement", ISO, Geneva, 1993; AOAC, 2002b; Codex Alimentarius, 2004; JCGM_100, 2008; EURACHEM / CITAC Guide CG 4, 2012)
Negative DNA target control	Reference DNA, or DNA extracted from a certified reference material, or known negative sample not containing the sequence under study. NOTE This control demonstrates that the results of analyses of test samples not containing the target sequence will be negative. (ISO 24276:2006)
Performance criteria	(27) requirements for a performance characteristic according to which it can be judged that the analytical method is fit for the intended use and generates reliable results. (Regulation (EU) 2021/808)
Positive DNA target control	Reference DNA, or DNA extracted from a certified reference material, or known positive sample representative of the sequence or organism under study. NOTE This control is used to demonstrate that the PCR reagents are working as intended. (ISO 24276:2006)
Precision	(28) the closeness of agreement between independent test results obtained under stipulated conditions and is expressed as the standard deviation or coefficient of variation of the test results. (Regulation (EU) 2021/808) The closeness of agreement between independent test results obtained under stipulated (predetermined) conditions. The measure of precision usually is expressed in terms of imprecision and computed as standard deviation of the test result. Less precision is determined by a larger standard deviation. (ISO 24276:2006; AOAC, 2016) An indication of a class of quality parameters. This class includes uncertainty, repeatability (accordance), reproducibility (concordance) and robustness, among some other additional parameters.
Probability of identification	Expected or observed fraction of test portions at a given concentration that gives positive result when tested at a given concentration. (AOAC, 2016)
Proficiency study	The analysis of one or more identical samples by a range of laboratories. The participants are allowed to choose their own methods, provided that these methods are used under routine conditions. The study is ideally performed according to ISO guide 17043 and can be used to assess the reproducibility of methods.
Qualitative method	(29) an analytical method, which detects or identifies a substance or a group of substances on the basis of its chemical, biological or physical properties. (Regulation (EU) 2021/808)
Quantitative method	(30) an analytical method, which determines the amount or mass fraction of a substance so that it may be expressed as a numerical value of appropriate units. (Regulation (EU) 2021/808)
Recovery	(31) the recovery corrected amount of an analyte divided by the fortified amount of the analyte in the matrix sample, expressed as a percentage. (Regulation (EU) 2021/808) Fraction or percentage of the analyte that is recovered when the test sample is analyzed using the entire method. There are two types of recovery: (1) Total recovery based on recovery of the native plus added analyte, and (2) marginal recovery based only on the added analyte (the native analyte is subtracted from both the numerator and denominator). (AOAC, 2016)
Reference material	(33) a material sufficiently homogeneous and stable with respect to one or more specified properties, which has been established to be fit for its intended use in a measurement process or in examination of nominal properties. (Regulation (EU) 2021/808)
Repeatability	(35) precision under conditions, where independent test results are obtained with the same method on identical test items in the same laboratory by the same operator using the same equipment within short intervals of time. (Regulation (EU) 2021/808) Precision under repeatability conditions. (ISO 24276:2006; AOAC, 2016)
Repeatability conditions	Conditions where independent test results are obtained with the same method on identical test items in the same laboratory by the same operator using the same equipment within short intervals of time. (ISO 24276:2006; AOAC, 2016)
Reproducibility	(36) precision under conditions, where test results are obtained with the same method on identical test items in different laboratories with different operators using different equipment. (Regulation (EU) 2021/808) precision under reproducibility conditions. (ISO 24276:2006; AOAC, 2016)
Reproducibility conditions	Conditions where test results are obtained with the same method on identical test items in different laboratories with different operators using different equipment. (ISO 24276:2006; AOAC, 2016)
Ruggedness (robustness)	(37) the susceptibility of an analytical method to changes in experimental conditions under which the method can be applied as presented or with specified minor modifications. (Regulation (EU) 2021/808)

Screening method	(38) means a method that is used for screening of a substance or class of substances at the level of interest. (Regulation (EU) 2021/808)
Screening target concentration	(39) the concentration lower than or equal to the CCβ at which a screening measurement categorises the sample as potentially non-compliant 'Screen Positive' and triggers a confirmatory testing. (Regulation (EU) 2021/808)
Selectivity	(40) the ability of a method to distinguish between the analyte being measured and other substances. (Regulation (EU) 2021/808)
Sensitivity	The frequency at which the presence of a target is confirmed in test samples; the share of true positives in the total number of examined spiked samples. Applied in qualitative visual methods. This parameter is part of the calculation of Correctness, together with Specificity. See Detection capability (CC*β*) for quantitative methods. Change in the response divided by the corresponding change in the concentration of a standard (calibration) curve. NOTE This is the slope of the analytical calibration curve. (ISO 24276:2006)
Single laboratory study (in-house validation)	(41) means an analytical study involving a single laboratory using one method to analyse the same or different test materials under different conditions over justified long time intervals. (Regulation (EU) 2021/808)
Specificity	The frequency at which the absence of a target is confirmed in test samples; the share of true negatives in the total number of examined blank samples. Applied in qualitative visual methods. This parameter is part of the calculation of Correctness, together with Sensitivity. The ability of a method to distinguish between the analyte being measured and other substances. This characteristic is predominantly a function of the measuring technique described, but can vary according to class of compound or matrix. Property of a method to respond exclusively to the characteristic or analyte under investigation. (ISO 24276:2006)
Target Taxon	Taxon to which the genetically modified organism belongs. (ISO 24276:2006) General: the taxon (species, type, group) to which the component found belongs. In visual examinations every unit (particle, etc.) found should in principle be confirmed as belonging to the target taxon. Targets are usually included and described in libraries used for identification. See Expert system.
Test portion	(45) means the quantity of material drawn from the sample on which the test or observation is carried out. (Regulation (EU) 2021/808)
Trueness	(46) the closeness of agreement between the average value obtained from a large series of test results and an accepted reference value. (Regulation (EU) 2021/808; ISO 24276:2006)
Unit (visual entity)	The principal elements (fragments, particles, globules, seeds, fruits) which are visible and belonging to the target ingredient, target contaminant or undesirable substance of a visual monitoring method.
Unit (SI system)	(47) those units described in ISO 80000 and Council Directive 80/181/EEC. (Regulation (EU) 2021/808)
Validation	(48) the demonstration by examination and the provision of effective evidence that the particular requirements of a specific intended use are fulfilled, through a single laboratory study or a collaborative study. (Regulation (EU) 2021/808)
Weight uncertainty	The diversity among mass fractions resulting from investigating replicate test portions by visual analysis. This parameter can be compared with measurement uncertainty for chemical quantitative methods.
Within-laboratory reproducibility (intermediate precision/in- house reproducibility)	(49) means measurement precision under a set of within-laboratory conditions in a specific laboratory. (Regulation (EU) 2021/808)

Appendix 3 Expert systems performance criteria

Expert or decision support systems exist in a variety of implementations or applications. In the current framework, the focus is on systems mimicking the observers' process of making a decision based on collected information. The decision is the final establishment of an identity of a specimen, i.e. seed, compound feed ingredient, bone fragment, pollen grain.

Expert systems consist of several parts. A library of targets (plants, animal particles, feed ingredients, types of starch, pollen species a.o.) should be available for browsing, providing descriptions, images and key features. The scope of the expert system and its library should be defined in an inclusivity table and a description of the excluded targets. The key element of an expert system is the interactive support of the process of identification by a process of entering defined observations. Usually a dichotomous key and/or a free-access key is provided. The result is either a final target matching the observations (dichotomous key) or a list of targets in matching order, with ideally one target showing a maximum similarity (100% match) with the observed entity (free-access key). If the expert system includes the option to save the entered observations together with the resulting documentation on identification, a collection of user documentations can be built for later use. Applications of this stored documentation can include confirmation of correct identification, quality assurance, training support, what-if runs for considering alternative observations, and exchange of documented identification runs among technicians. WFSR has developed a platform for developing expert systems intended to support identification issues in visual feed and food safety research (Determinator: van Raamsdonk et al., 2010; 2012d; Uiterwijk et al., 2013).

A classification system for rice in the framework of the Combined Nomenclature (Regulation (EU) 1308/2013, Annex II, Part I) has been used in a proficiency study. The structure of the proficiency test was chosen to provide information for the validation of the classification model, implemented in the platform Determinator (CLEN, 2018).

Options for validation of the classification model containing the logic rules for running the keys should be available. There are three options in order of application:

1. The relationship between the targets in the library and the specified character states (the logic rules) needs to be confirmed. This is primarily a task of an expert in the domain of the expert system. In an ideal situation this expert was not involved in the development of the system.
2. An expert system should be capable to identify every target uniquely. In other words, no overlap among targets in the diversity space of the system should occur. This requirement assures that with correct observations the result cannot include two 100% matches with two different targets. A procedure for calculating this criterion is included in the platform Determinator (van Raamsdonk et al., 2012).
3. Reproducibility of the identification or classification of investigated material. These samples might have been identified in the past by an expert and be included in a reference collection. The original identity should not be achieved by the expert system to be validated. Correct application of the system should result in identical identifications.

Performance criteria have been developed for the design of expert systems (van Raamsdonk et al., 2012d). The basic framework is to consider a classification system of *m* targets with *n* features as a diversity space with *n* dimensions. Every target possesses a part of that diversity space. Objectives in designing a classification model are the absence of overlap of two or more target areas, and minimising the space between the targets.

The performance criteria include:

- Pearson correlation among features: the correlation matrix of features should not contain values exceeding an absolute threshold. The level of this threshold can be chosen depending on the object. Features with a very high correlation are representing largely the same information, resulting in an overrepresentation of that information.

- Redundancy: overlap between the areas of two targets would result in a probability to have two 100% matches of the subject with two different targets in an identification run. Overlap can be calculated for each pair of targets or for all targets in the classification model.
- Uniqueness: at least one feature should be able to distinguish every pair of two targets in the system. This criterion needs to be TRUE for all pairs of targets. In the situation that only one feature discriminates between two targets in a system, an erroneous observation for that feature would result in a false or confusing identification. Therefore, at least two discriminating features for every pair of targets should be included in the classification model.
- Separation capability of the classification model: the basic objective is a 100% separation capability. In specific cases a set of observations might point to two targets in the classification model (i.e. uniqueness is FALSE for at least one pair of targets), followed by further discrimination based on additional data from the library.

Appendix 4 Overview of scope of Guidelines

Guideline	Domain	Qualitative	Quantitative	Single or interlab	Criteria	Workflow	Comment
ISO/IEC 17025:2017. General requirements for the competence of testing and calibration laboratories.	general	no	yes		range, accuracy, measurement uncertainty, LOD, LOQ, selectivity, linearity, repeatability, reproducibility, robustness, bias	Sampling, control samples, reference material	General aspects of quality assurance
Analytical chemistry							
European Union							
European Union, 2021. Commission Implementing Regulation (EU) 2021/808 of 22 March 2021 on the performance of analytical methods for residues of pharmacologically active substances used in food-producing animals and on the interpretation of results as well as on the methods to be used for sampling. OJ L 180, 21.5.2021, p. 84-114.	chem	few	yes	single	Accuracy, precision, recovery, selectivity, specificity, CCα, CCβ, ruggedness, repeatability, reproducibility, measurement uncertainty	Control samples, certified or other reference material	
DG-SANCO, 2009. Guideline for initial validation and transfer of screening methods. Non-Paper 08/07/2009.	chem, microbiol	yes	yes	single	CCβ, action limit, cut-off level, specificity, rugg		2.8: SOP drafted before validation, 5.1.1: >59 samples per target/matrix combination for qualitative tests
ISO							
ISO, 1994a. Accuracy (trueness and precision) of measurement methods and results. Part 1. General principles and definitions. ISO 5725-1:1994(E).	chem	no	yes	inter: prof	Repeatability, reproducibility, bias		
ISO, 1994b. Accuracy (trueness and precision) of measurement methods and results. Part 2. Basic method for the determination of repeatability and reproducibility of a standard measurement method. ISO 5725-2:1994(E).	chem	no	yes	inter: prof	Repeatability, reproducibility		
ISO, 1994c. ISO 5725-4:1994(E).	chem	no	yes		Trueness		

Guideline	Domain	Qualitative	Quantitative	Single or interlab	Criteria	Workflow	Comment
ISO, 2010. ISO/IEC 17043: 2010. Conformity assessment - General requirements for proficiency testing. CEN/CENELEC, Brussels: Ref. No. EN ISO/IEC 17043:2010 E	chem	no	yes	inter: prof	Trueness, uncertainty, Z-scores, zeta-scores, En-scores; no parameters for qualitative tests		4.5.1: consistent with participants' routine procedures; Outliers
ISO, 2015. Statistical methods for use in proficiency testing by interlaboratory comparisons. International Standard Organisation, Geneva: ISO/IEC 13528:2015(E).	chem	yes	yes	inter: prof	Measurement uncertainty, bias, ranks, %rank, Z-scores, zeta-scores, En-scores		Robust statistics; 11.4.3: Gower coefficient
IUPAC							
Horwitz, W., 1995. Protocol for the design, conduct and interpretation of method-performance studies (IUPAC guidelines). Pure & Appl. Chem. 67: 331-343.	chem	no	yes	inter: val	Repeatability, reproducibility, bias, recovery, applicability		2.2: entrance test; 3.1: valid data; 3.4: outliers; 4.0: outlying laboratories
Currie, L.A., 1995. Nomenclature in evaluation of analytical methods including detection and quantification capabilities (IUPAC Recommendations 1995). Pure & Appl. Chem., Vol. 67, No. 10, pp. 1699-1723.	chem	no	yes	single, few interlab	Calibration function, sensitivity, evaluation function, LOD, LOQ, CCα, CCβ	Certified (or "Standard") reference materials	
Thompson, M., R. Wood, 1995. Harmonized guidelines for internal quality control in analytical chemistry laboratories. Pure & Appl. Chem., Vol. 67, No. 4, pp. 649-666.	chem	no	yes	single	Measurement uncertainty	Duplicate test materials, control samples, reference materials. Uncertainty is an integral part of the ICQ.	Quality control of sampling, in-line analyses, multivariate methods, calibration are excluded.
Thompson, M., S.L.R. Ellison, R. Wood, 2002. Harmonized guidelines for single laboratory validation of methods of analysis (IUPAC Technical Report). Pure Appl. Chem., Vol. 74, No. 5, pp. 835–855.	chem	no	yes	single	Applicability, selectivity, linearity, trueness, precision, recovery, range, LOD, LOQ, sensitivity		
AOAC							
AOAC, 2002b. Guidelines for Single Laboratory Validation of Chemical Methods for Dietary Supplements and Botanicals.	chem	no	yes	single	Applicability, selectivity, calibration, trueness, recovery, repeatability, measurement uncertainty, reproducibility, LOD		3.4.2: RSDr series: RSDr = C – 0.15 with C = concentration expressed as a mass fraction; 3.4.4: RSDR series: RSDR = 2·C – 0.15 with C = concentration expressed as a mass fraction

Guideline	Domain	Qualitative	Quantitative	Single or interlab	Criteria	Workflow	Comment
AOAC, 2002a. Appendix D: Guidelines for collaborative study procedures to validate characteristics of a method of analysis.	chem	no	yes	inter: val	Bias, recovery, trueness, repeatability, reproducibility		In part identical to IUPAC Guideline collaborative validation studies (Horwitz, 1995). 1.7: familiarisation with samples towards satisfactory performance; 2.2: final selection of participants based on capabilities and past performance; 4.2: method should be followed exactly; 5.2: outliers
AOAC, 2016. Appendix F: Guidelines for Standard Method Performance Requirements.	chem?	yes	yes	single	Quantitative: applicability bias, precision, recovery, LOQ, reproducibility, measurement uncertainty Qualitative: inclusivity/selectivity, exclusivity/cross-reactivity, environmental interference, lab variance, probability of detection		Table A6, Annex D: RSDr series: Predicted RSDr = C – 0.15 with C = concentration expressed as a mass fraction
Codex Alimentarius							
Codex Alimentarius, 2008. Guidelines for the validation of food safety control measures. CAC/GL 69 – 2008.	chem, microbiol	(no)	yes	single			In-line metal detector for metal particles
Codex Alimentarius, 2004. Guidelines on measurement uncertainty. CAC/GL 54-2004.	chem	no	yes				Measurement uncertainty. RSDR series: LL= 0.56 C, UL= 1.44 C, with C = concentration expressed as a mass fraction
Other							
Pocklington, W.D., 1990. Harmonized protocols for the adoption of standardized analytical methods and for the presentation of their performance characteristics. Pure & Appl. Chern., Vol. 62, No. 1, pp. 149-162.	chem	no	yes	single	Precision, trueness, repeatability, reproducibility, sensitivity, LOD		I-1): RSDR series: RSDR = 2(1 – 0.5·log·C) with c = concentration expressed as a decimal fraction; II-1): outliers
JCGM_100_2008. Evaluation of measurement data — Guide to the expression of uncertainty in measurement.	chem/phys	no	yes	single	measurement uncertainty		

Guideline	Domain	Qualitative	Quantitative	Single or interlab	Criteria	Workflow	Comment
Microbiology							
ISO, 2003. Microbiology of food and animal feeding stuffs – Protocol for the validation of alternative method. ISO/FDIS 16140:2003(E).	microb						
Feldsine, P., C. Abeyta, W.H. Andrews, 2002. Validation of Qualitative and Quantitative Food Microbiological Official Methods of Analysis. J AOAC, 85: 1187-1200.	microb	yes	yes				
EPA, 2016. Microbiological Methods of Analysis, prepared by The FEM Microbiology Action Team. FEM Document Number 2009-01, REVISION: December 21, 2016.	microb	yes	yes	single, inter: val	Specificity, sensitivity, repeatability, reproducibility, recovery, bias, LOD, linearity, calibration		2.6.5: Margin of error related to number of laboratories and sensitivity; requirements for experience and expertise of participants
FDA, 2015. US Food & Drug Administration Office of Foods and Veterinary Medicine.	microb	yes	yes	single, inter: val	Only named in Glossary: accuracy, applicability LOD, LOQ, linearity, precision, recovery, repeatability, ruggedness, specificity, sensitivity, systematic error, trueness, uncertainty		5.0: comparison among methods by t-test for accuracy and F-test for precision
Molecular biology							
ISO 24276:2006. Foodstuffs -- Methods of analysis for the detection of genetically modified organisms and derived products -- General requirements and definitions. (reviewed and confirmed in 2015)	mol biol			single	LOD, LOQ	General requirements for facilities, sampling, negative and positive controls, environmental control	
ISO 20813:2019. Molecular biomarker analysis -- Methods of analysis for the detection and identification of animal species in foods and food products (nucleic acid-based methods) -- General requirements and definitions	mol biol	yes	yes	single, inter: prof	Specificity (inclusivity, exclusivity), sensitivity, LOD, LOQ, precision, trueness, robustness, probability of detection	General requirements for facilities, sampling, negative and positive controls, environmental control	

PART 2 Validation

1 Background

Monitoring of biological and physical hazards, as part of the BCMP cocktail of hazards for feed and food production (biological, chemical, microbiological, physical hazards), is partly based on visual observations, including but not limited to microscopy. In terms of actual monitoring, visual methods include targets or contaminants such as bone fragments or other particles of animal origin, plant seeds, spore bodies of moulds (sclerotia), packaging material, and "Besatz". Identification of legal ingredients and composition analysis is part of the domain of visual examination as well.

Methods for visual or microscopic examination include other assumptions than those used in analytical chemical methods. This has four reasons:

1. Inhomogeneity, due to a low number of large units.
2. The size of the analysis sample is typically in the range of grams, from 10 grams for analysis of animal proteins (microscopic) up to 500 grams for undesirable substances (macroscopic).
3. The sensitivity of visual methods is in the ppm (mg/kg) range or higher (25 ppm = 0.0025%). Levels in the ppb range can be assumed to be below the detection level.
4. There is a principal difference between the protocol and the examination. A protocol can and should be harmonised. The detection and identification are performed by the microscopic technician based on knowledge and experience. In all cases the performance of the method and the performance of the technician has to be addressed separately.

In a range of cases applications of quality parameters for visual methods are adopted from chemical standards. In specific cases such as qualitative monitoring methods, other principles should be explored for defining quality parameters in the domain of visual research.

1.1 Scope of the Guidance

This Guidance presents an elementary overview of the application of visual research. It evaluates different quality parameters and criteria taken from other major disciplines such as analytical chemistry and microbiology. The specific application of relevant parameters to visual detection methods including microscopy will be discussed and a framework of dedicated sets of quality parameters for the domain of visual monitoring methods will be given, separate for quantitative, qualitative and identification methods. Elements of visual methods are presented and discussed in relation to quality parameters.

Part 1 has provided the theoretical background and principles of quality assurance and control for visual inspection. Definitions and explanations of quality parameters are included in this part as well. The current Part 2 consists of an overview of the relevant parameters and procedures for application to visual examination methods. Short introductions will be provided for the parameters included. However, further background with literature references and applicable confidence limits are given in Part 1.

2 Domain of visual methods for feed and food

The set of methods of the domain of visual inspection covers a range of magnifications of the target under study. The scope of visual methods includes anything that is or can be made visible (or the absence of something). The domain includes three levels of particle size, each demanding its own approach and equipment (Table 1).

Table 1 *Overview of three levels of particle size with the appropriate equipment and examples of application.*

Equipment and magnification	Type of matrix	Particle size	Type of method
Visual, non-microscopic: none or magnifying glass; 1 to 8x	Unground or raw materials	Preferably larger than 1 mm	Quantitative
Macroscopic: binocular or stereo microscope; 8 to 64x	Ground materials; coarse	200 – 1000 µm	Semi-quantitative
Microscopic: compound microscope; 100 – 400x	Powder or meal, fluids with cells or particles; fine	10 – 200 µm	Qualitative

3 Quality parameters

In a series of occasions quality standards differentiate among quantitative and qualitative methods for the detection of a contaminant or an undesirable substance. Quality parameters apply which are partly comparable, some have a different statistic background, and some are not relevant. Table 2 provides an overview of quality parameters for the mentioned two types of methods. Several other types of methods are being used in the domain of visual inspection which are not aiming at the detection of a specific component. These include establishing the identity of a matrix or a specimen, and the estimation of a composition. The latter can be considered to be the establishment of a series of two or more identities, accompanied with an estimation of their shares in the total sample material. These special types of methods will be treated separately.

Table 2 *Overview of quality parameters for visual inspection, organised in three different types of methods.*

	Accuracy			Precision			
Parameter	Recovery	Limit	Specificity	Uncertainty	Repeatability	Reproducibility	Robustness
Quantitative	yes[1]	yes	yes	yes[1]	(no)[2]	yes[1]	yes
Parameter	Correctness: sens+spec	Sensitivity	Selectivity	U	Accordance	Concordance	Robustness
Qualitative	yes	yes	yes	no	yes	yes	yes
Parameter	Trueness					Reproducibility	
Identification	(yes)[3]					yes	

1: with an adapted background and a modified information content.

2: might be applicable in specific circumstances.

3: only after analysis of reference material with approved identity.

Basic statistic theory can be found in handbooks such as Sheskin (2002) and Agresti et al. (2018).[3]

The experimental designs are presented in boxes with red font.

The following overview provides practical guidance for experimental design, calculation and interpretation of the parameters.

3.1 Quantitative methods

In those cases where material can be filtered or selected manually, quantitative parameters can be calculated from the results of the sample inspection. The nature of the contaminant will determine the type of data. Target material will be selected as single units for seeds (e.g. *Ambrosia*, *Datura*) or ergot sclerotia. These units can be counted and weighted. Although targets such as remnants of packaging material or Besatz principally consist of units as well, counting is usually not practised. The smaller and, hence, the more numerous the units are, the smaller the problem of inhomogeneity among subsamples might be. Manual selection is generally not feasible for targets of which the units are generally smaller than 1 mm. This size limit applies to e.g. particles of animal proteins or smaller microplastic, for which the results are usually expressed as counts per amount of matrix material.

Inhomogeneity is a factor in visual inspection with a larger impact than occurring in other analytical disciplines. As a consequence, duplicate samples and analyses are usually not achievable. Samples should be spiked individually, and an a-priori known spike level of a reference or aggregate laboratory sample is not

[3] All references to literature have been included in Part 1.

available. Repeatability cannot be calculated in most cases and several other parameters need a modified strategy for calculation.

At least eight samples per spike level are spiked with a number of units at or below the legal limit. Since units can vary in weight, either within reasonable limits (seeds) or considerable (ergot sclerotia), the chosen amount of material per spike level can be adjusted in terms of numbers or in terms of weight. The latter is recommended since final results will be related to legal limits, which are expressed as weight percentages. Production of one batch of a laboratory sample and division in eight analysis samples will result in largely different spike levels, due to inhomogeneity. It is necessary to select material of the contaminant with the approximate intended amount for representing the chosen spike level, and add this to an amount of blank matrix material. Minimum sample size per analysis sample is 500 grams. As alternative, eight samples from practice can be used to collect data for calculating the parameters for precision. The advantage is to have a variable range of, initially unknown, contamination levels mimicking real life diversity.

For Accuracy parameters, spike amounts per sample per level need to be known. For Precision parameters, every sample, either spiked or derived from practice, needs to be analysed twice, meaning that the selected units are mixed in the sample material again after counting and weighting, ready for a second analysis. In addition, samples which are additionally spiked with a mimicking type of units and samples inspected under modified circumstances are to be analysed for the quality parameters specificity and robustness. The samples are to be analysed under intralaboratory reproducibility conditions. In the following equations n equals 8 or more and k equals 2 or more. The detection of *Ambrosia* seeds in bird feeds is presented in Appendix 1 as example for this type of methods.

3.1.1 Recovery/bias

At least 8 blank samples with a minimum of 500 gram each: n; store the sample weights w_i
↳ spike each sample i individually at a known spike level: s_i
↳ select spike material from each sample i: r_i
Perform this experiment at least at two spike levels, one close to the legal limit, and a second close to the expected level of detection

Average Recovery, expressed as percentage, is calculated as

$$R\% = \frac{\sum_1^n (r_i/s_i)}{n} * 100 \qquad (1)$$

with r_i as the number or amount of recovered material, and s_i as the number or amount of spiked material for sample i. The recovery can be based on either the counts or the weights (w/w) and is expressed as percentage. This equation deviates from the usually applied procedure, which calculates the quotient of the average recovered material divided by the sample mean. This procedure cannot be used for individually spiked samples. The spike levels for calculating the recovery can be chosen at a reasonable level close to the expected detection limit (one or two units per kg), and close to the legal limit. A level close to the analytical threshold, usually between 40-60% of a legal limit, can be chosen as well.

It might be necessary in certain cases to calculate the standard deviation of the Recovery. In the absence of a batch with an average contamination level, correction for the individual spike level is necessary. The SD of the Recovery (in percentages) can be calculated as

$$SD_R = \sqrt{\sum_1^n (t_i - R\%)^2 \Big/ (n-1)} \qquad (2)$$

with $t_i = r_i/s_i * 100$. The average of the deviation between the values t_i and $R\%$ (Equation (1)) can be calculated in a numerically correct way, but need further explanation. Results representing different spike levels can be included in one result for Recovery, since a correction for differing spike levels is included in the equation. Spike levels with the intention to be different (e.g. low and high contamination) can be used for calculating one result for *R%*, but could largely influence the standard deviation of *R%* (equation (2)). This is not the intention as indicated in the experimental box, and separate values for *R%* and SD_R should be calculated for every spike level.

Bias is the still existing systematic deviation under reproducibility circumstances of the recovered amounts from the samples' spike levels expressed in mg/kg:

$$\delta = \frac{\sum_1^n |r_i - s_i|}{n * \overline{w}} \quad (3)$$

with r_i and s_i as the recovered and spiked levels per sample i, and $\overline{w}$ as the average weight of the analysis samples. In optimal reproducibility circumstances and a well-established specificity the systematic bias of the counts would be zero. The connected bias for weight percentages (w/w) could be influenced by uncontrolled parameters.

3.1.2 Detection limit CCβ: quantification limit

The results of the experiment as set out in Paragraph 3.1.1. with the lowest spike level are being used for the calculation of the standard deviation of the measured values of spiked samples:

$$SD_{DL} = \sqrt{\sum_1^n (r_i - s_i)^2 \Big/ (n-1)} \quad (4)$$

with r_i, s_i and n as used in equation (1). The detection limit is defined as three times this standard deviation and expressed as mg/kg:

$$DL = 3 * SD_{DL} \Big/ \overline{w} \quad (5)$$

with SD_{DL} taken from equation (4) and $\overline{w}$ as the average weight of the analysis samples.

3.1.3 Selectivity/specificity

A. At least 2 blank samples with a minimum of 500 gram each: k
  ↳ spike each sample i individually at a known spike level: s_i; add mimicking material at a known level
    ↳ select spike (target) material from each sample i: r_i
  Repeat this experiment when desired with different types of mimicking materials

Alternatively:

B. At least 2 blank samples with a minimum of 500 gram each: k
  ↳ intended target is not added: $s_i = 0$; add mimicking material at a known level
    ↳ select spike (target) material from each sample i: r_i

The precise selection of the units belonging to the target material among other mimicking units needs to be confirmed. Therefore, at least two samples spiked with material of the intended target together with material of a look-alike ingredient will be analysed. More than one look-alike ingredient may be included in this experiment that should be checked. Examples are: different seeds in case of *Ambrosia*, hard sweets, pasta and chocolate, in case of packaging material. Also matrix effects may influence the selection of the target (example: all seeds of a bird feed of the same size in case of *Ambrosia* preventing concentration by sieving). If more than one factor of mimicking particles or a matrix effect is chosen, only one factor per sample need to be applied.

Recovery in the presence of a mimicking material or matrix effect can be calculated as

$$R^* = \frac{\sum_1^k (r_i^* / s_i^*)}{k} * 100 \quad (6)$$

with r_i^* as the number or amount of recovered material, and s_i^* as the number or amount of spiked material for sample i in the presence of a mimicking ingredient or a modified matrix. In the situation that one or more blank samples have been contaminated with only mimicking (non-target) material, the recovery cannot be calculated in the situation that the denominator per sample is zero ($s_i = 0$). Equation (7) can be used for calculating the performance of the method based on blank samples.
The relative deviation for specificity will be calculated as follows, expressed in mg/kg:

$$\delta^*_s = \frac{\sum_1^n |r_i - s_i|}{k * \bar{w}} \quad (7)$$

with r_i^* and s_i^* as defined for equation (6).

3.1.4 Repeatability and reproducibility

A complete calculation of both repeatability and reproducibility standard deviations requires the availability of three subsamples for analysis, taken either from: - a laboratory sample, - a spiked blank, - a reference sample or a sample from practice, which are in all cases sufficiently homogenised. Two of these subsamples are to be analysed in the same run as duplicates under repeatability conditions. The third sample is to be analysed later under reproducibility conditions. Samples subjected to visual examination usually suffer from two conditions due to the principal presence as units: (a) at least 500 gram of material is required per analysis of one subsample, and (b) a sufficient homogenisation of the laboratory sample depends on the size of the targeted units and the type of the matrix (Part 1, section 4.1.5). The ideal starting situation per laboratory sample is the availability of 1500 gram sufficiently homogenised material. In the absence of one or both premises (amount and homogeneity), an alternative is to reintroduce the selected portion of the target material in the remaining matrix material after the first analysis and analyse the same sample for the second time under reproducibility conditions. This approach fails for destructive methods.

The approaches for calculating the standard deviations for repeatability, divisibility and/or reproducibility are summarised in Table 3.

Table 3 *Approaches for the calculation of repeatability and reproducibility in combinations of several analytical circumstances.*

Type of sample: Type of method:	Homogenised, at least 1500 gram total sample material: subsampling can be applied		Not homogenised, at least 1500 gram total sample material: no true subsamples		Not homogenised sample material: reintroduction of target material	
Non-destructive (matrix and target remains unchanged, if required the target can be reintroduced for a second analysis)	A	Repeatability: s_r Reproducibility: s_{Rw}	B	Repeatability: s_r Reproducibility: s_{Rw}	C	Reproducibility: s^*_{Rw}
Destructive (matrix is damaged, diluted or destroyed, target cannot be replaced for a second analysis)	A	Repeatability: s_r Reproducibility: s_{Rw}	B	Repeatability: s_r Reproducibility: s_{Rw}	D	None

Situations A and B: a batch of material per sample of at least 1500 gram is available. The numerical calculations for situation **A** and **B** are identical regardless of the state of homogenisation. Any type of material, suitable for macroscopic evaluation can be subjected to this approach. For situation **A**: (semi-) fluids, fine granular matrices (packaging material in candy syrup, plastic in soil or manure), for situation **B**: whole kernel feeds (cereals, bird feeds). Some matrices can be diluted by an amount of extra fluid, which should be a known quantity in order to allow the calculation of the original contamination level (known mass balance). This is an example of a destructive pre-treatment.

At least 8 samples from practice with unknown levels of the target higher than zero OR at least 8 blank samples spiked at different levels, with a minimum of 1500 gram each: n
- ↳ homogenise as good as possible every sample i and produce three subsamples
  - ↳ examine subsample *1* and *2* of each sample i individually and select target material from each sample i: r_{i1} and r_{i2}
    - ↳ wait at least a couple of days to reach reproducibility circumstances
      - ↳ examine subsample *3* of each sample i individually and select target material from each sample i: r_{i3}

The repeatability standard deviation will be calculated from the results of the duplicate analysis, obtained under repeatability conditions:

$$s_r = \sqrt{\sum_1^n (r_{i1} - r_{i2})^2 / 2n} \quad (8)$$

with r_{i1} and r_{i2} as the recovered amounts from the two replicates expressed in mg/kg. The standard deviation under divisibility circumstances s_d will be calculated from the same type of primary results. The average of these two amounts can be calculated as m_i. Within laboratory reproducibility standard deviation can be calculated as:

$$s_{Rw} = \sqrt{\left(\sum_1^n (r_{i3} - m_i)^2 / 2n\right) + \frac{1}{4} * s_r^{\,2}} \quad (9)$$

with r_{i3} as the recovered amount of the third analysis sample analysed under reproducibility conditions and m_i as the average of the first two determinations, both values expressed in mg/kg. The availability of material with sufficient amount and sufficient quality for visual inspection is unlikely to happen in most cases. If the values for s_r and/or for s_{Rw} do not meet the a-priori set conditions, an a-posteriori conclusion that apparently the sample material was not sufficiently homogenised cannot be drawn, and an a-posteriori conclusion that apparently situation **B** applied instead of situation **A** cannot be drawn. Failing to meet the a-priori set conditions could also indicate that the method is indeed not repeatable and/or reproducible. Therefore, the strategy **A** can only be applied in the case that the materials to be studied are a-priori approved to meet all the conditions for repeatability necessary for the intended analysis. The values for s_r and for s_{Rw} in situation **B** will be (much) larger than generally found after chemical analyses, which is also the domain where these parameters are frequently applied. It is therefore recommended to add explanations of the background when presenting validation results.

Situation C: a homogenised batch is not available or cannot be produced per sample. General type of material as indicated for options **A** and **B**. Prerequisite for situation **C** is a non-destructive method. A measure for reproducibility which is calculated after a procedure of reintroduction, indicated by s^*, needs a specific interpretation.

At least 8 samples from practice with unknown levels of the target higher than zero OR at least 8 blank samples spiked at different levels, with a minimum of 500 gram each: n

↳ examine each sample individually and select target material from each sample i: r_{i1}

↳ reintroduce the selected target material in each sample i; homogenise and store for a couple of days up to one week to regain a levelled moisture and fat content

↳ examine each sample individually for the second time and select target material from each sample i: r_{i4}

NOTE: This experiment can be combined with the experiment in paragraph 3.1.1 when executed in duplicate, as described above

In the absence of duplicate samples and results and, hence, the lack of a value for the repeatability standard deviation, the calculation of the reproducibility standard deviation needs an alternative approach. Therefore, calculation of the reproducibility standard deviation will be based on the difference between the pair of analyses per sample (second analysis carried out after replacement of the material selected in the first analysis):

$$s^*_{Rw} = \sqrt{\sum_1^n (r_{i1} - r_{i4})^2 / 2n} \quad (10)$$

with r_{i1} and r_{i4} as results of the two analyses of sample i under reproducibility conditions. For the current type of methods these conditions include at least one week difference between the two analyses and examination preferably by two different technicians for each sample. In the situation of a combined experiment for recovery (3.1.1) and reproducibility (3.1.4), the spike levels of the 8 or more samples need to be (highly) comparable. The value of S_{Rw} will then only provide information for the chosen spike level. Additional experiments can be carried out for other contamination levels.

It is essential that correct a-priori limits are fixed for the parameters repeatability and reproducibility. For whole kernel matrices and undesirable substances of the same size (situation **B**) these values can be extracted from the Horwitz equation in section 4.1.4 of Part 1. In situations where a better homogenisation can be achieved lower limits for these two parameters will apply, but these values might still be higher than usually achieved in chemical analysis. An example of an intermediate situation and approaches to fix a-priori upper limits is presented in the WFSR validation report of the detection method for packaging material in candy syrup.

Expression of relative repeatability and reproducibility

It can be desired to express repeatability and reproducibility relative to the spike level of the samples in the dataset. The usual way to calculate the relative standard deviation (RSD) or the coefficient of variation (CV) is the equation

$$RSD = s/\mu, \quad (11)$$

with s as the repeatability or reproducibility standard deviation, and μ as arithmetic mean. This approach can be followed in all situations, but the interpretation is different. In situations **A** and **B** the value for RSD represents an uncertainty including at least two factors: the inhomogeneity among the subsamples of a sample and the expertise level of the technician. In situation **C** inhomogeneity is principally no factor and the capacity of the technician to follow the procedure and interpret the selected material is the primary cause of variation, besides minor factors such as air moisture and temperature during examination.

Combination of approaches

The different approaches for calculating one or more parameters depend on analytical circumstances: the amount of material available and the nature of the method. The classification of these approaches result in overlapping strategies. If a non-destructive method is applied in the presence of an amount sufficient for subsampling (Table 3 top row, option **A** and **B**), the selected targeted material can be reintroduced in one or more subsamples and a second examination can be carried out (Table 3 top row, option **C**).

Assume a sample with an undesirable substance, which can be divided in three subsamples.

At least 8 samples from practice with unknown levels of the target higher than zero OR at least 8 blank samples spiked at different levels, with a minimum of 1500 gram each: n

↳ homogenise as good as possible every sample i and produce three subsamples

↳ examine subsample *1* and *2* of each sample i individually and select target material from each sample i: r_{i1} and r_{i2}

↳ wait at least a couple of days to reach reproducibility circumstances

↳ examine subsample *3* of each sample i individually and select target material from each sample i: r_{i3}

↳ reintroduce the selected target material in the first subsample of each sample i; homogenise and store for a couple of days up to one week to regain a levelled moisture and fat content

↳ examine the recontaminated subsample of each sample i individually for the second time and select target material from each sample i: r_{i4}

Three parameters can be calculated:
- Calculation of repeatability standard deviation s_r from r_{i1} and r_{i2} with equation 8.
- Calculation of reproducibility standard deviation s_{Rw} from m_i and r_{i3} based on equation 9, with m_i as average of the results r_{i1} and r_{i2} for sample i.
- Calculation of reproducibility standard deviation s^*_{Rw} from r_{i1} and r_{i4} based on equation 10.

Note that this strategy includes different versions of a reproducibility standard deviation. The standard deviation after reintroduction can be expected to be low, close to zero. At the other end of the spectrum, the reproducibility standard deviation among subsamples of a non-homogeneous sample would be considerable. The following order of the level of the parameters can theoretically be assumed:

$$s^*_{Rw} < s_r < s_{Rw}$$

3.1.5 Uncertainty: count dispersal and weight uncertainty

The calculation of a parameter for uncertainty is only necessary in the situation that the examined samples are extracted from a homogenised batch of sample material, subsamples from one (aggregate) sample (situation **A** in Paragraph 3.1.4). Parameters for uncertainty can be calculated from the counts of the units used for spiking or from their weights. The description of the type of samples classified under situation **A** implies that counts will not be expected to be available.

Uncertainty can be expressed in different notations. Here, the relative versions of the standard deviation and the bias are being used (for motivation: Part 1, section 4.1.4). The calculation of the uncertainty is based on the reproducibility standard deviation (equation 9) and the bias under reproducibility circumstances (equation 3), expressed as relative values calculated according to equation 11:

$$u = \sqrt{\left(\frac{s_{Rw}}{\mu}\right)^2 + \left(\frac{\delta}{\mu}\right)^2} \quad , \tag{12}$$

and the expanded uncertainty:

$$U^* = 2 * u \tag{13}$$

In the case of using samples from practice the bias δ is not available and the uncertainty would solely be based on the reproducibility standard deviation (equation 9).

The applicability of the expanded uncertainty U^* depends on the type of target material. Considering methods capable of producing both counts and weights (weed seeds, ergot sclerotia) applied under reproducibility circumstances, the dispersal of the *counts* for these methods is expected to be low. This parameter could be used as measure for process control. The expanded uncertainty based on the *weights* would show a larger value, since a higher number of environmental factors will influence the result. This larger uncertainty is relevant for the reported result. The existence of two values for dispersal or uncertainty will not hold for other types of targets. Results for the inspection of packaging material and comparable methods will only provide documentation for uncertainty in terms of weight (w/w) or percentages.

Data from experiments with *Ambrosia* seeds, *Datura* seeds and ergot sclerotia in bird feeds and whole kernel cereals revealed that the relationship between contamination level and the relative standard deviation under reproducibility circumstances can be described as:

$$RSD_{Rw} = C^{-0.41} \quad (14)$$

With RSD as measure for weight uncertainty in the absence of a value for bias. The expanded weight uncertainty will follow:

$$U^* = 2 \cdot C^{-0.41} \quad (15)$$

The derivation of these relationships between contamination level and uncertainty is discussed in Appendix 1 in Part 1 of this Guidance. An overview of uncertainty limits in the situation of a granular undesired substance in a granular matrix (weed seeds, ergot sclerotia) is presented in Table 4.

Table 4 *Maximum values for uncertainty of results expressed as weight % (w/w) for undesired substances in dry granular material after subsampling assuming an average unit length of 5 mm, related to the level of contamination.*

Spike	Concentration mg/kg	Expected RSD_R	Expanded uncertainty (w/w)
0.0025%	25	77	not applicable
0.005%	50	58	not applicable
0.01%	100	44	87
0.02%	200	33	66
0.05%	500	23	45
0.1%	1000	17	34

3.1.6 Robustness/stability

At least 2 blank samples with a minimum of 500 gram each: k
↳ spike each sample i individually at a known spike level: s_i
↳ select spike material from each sample i under the circumstance of one modified factor of the original method: r_i
Repeat this experiment when desired with modification of another factor, one in each experiment

A method can be considered robust under the condition that it is implemented and applied as originally designed. Specific parameters such as application of specific sieves in methods for granular matrices, the time needed to properly monitor the entire sample material, the temperature used to extract the target from the matrix or to inactivate the matrix are examples of conditions which can be modified for testing the robustness of a method. One or more parameters can be modified for testing this robustness, but only one has to be chosen per experiment.

The relative deviation per modified parameter of the method can be calculated according to the following equation, expressed in mg/kg:

$$\delta^*_{m1} = \frac{\sum_1^n |r_{1i} - s_{1i}|}{k * \bar{w}} \quad (16)$$

with r_{1i} as the number or amount of recovered material, and s_{1i} as the number or amount of spiked material for sample i for modification 1. Subsequent experiments, each directed to another parameter, can be carried out independently with order numbers 1, 2 etc. A part of the bias found in these experiments points to the general methodological bias as calculated from the results in paragraph 3.1.1 (Equation (3)). In order to find the bias which can be specifically attributed to the modified factor in the experiments in this paragraph, the following calculation can be followed:

$$\Delta\delta_{m1} = \delta^*_{m1} - \delta \quad (17)$$

If recovery and relative deviation calculated from experiments looking for effects of modifications exceeds the appropriate limits, a method can still be declared validated. In those cases, the conclusion is that the

investigated parameters and circumstances are critical (with indication of the boundaries) for a reliable application of the method.

3.2 Qualitative methods

Results will be collected at the level of (sub-)samples: absence or presence of the target. This requires larger numbers of samples to be analysed in order to retrieve reliable statistics. Nevertheless, initial validation should be carried out in one laboratory instead of using data from proficiency tests in order to avoid between-laboratory differences of expertise. When considering the need of at least 95% of correct results, the minimum number of analyses per parameter and per spike level should be 20.

Spiking by means of stepwise dilution for producing a batch per treatment can be applied for spike levels down to 0.05%, as can be concluded from a range of organised interlaboratory studies for the detection of processed animal proteins. This approach is not documented for other contaminants. Samples with spike levels below 0.05% are to be produced by individual spiking. Sufficient homogenisation of the sample material immediately before starting the analysis is necessary for those low levels. The detection of processed animal proteins in compound feeds is presented in Appendix 1 as example for this type of methods.

3.2.1 Correctness (accuracy)

A. Produce a homogenised batch of sample material with the target at the intended spike level, with an amount sufficient for taking the planned number of samples
At least 8 (n) laboratory samples with the minimum required amount of material are taken and analysed in duplicate; the number of results is at least 16 ($2*n$)
↳ analyse the presence or absence of the target in the samples: TP and FN
Perform this experiment at least at two spike levels, one close to the legal limit, and a second close to the expected level of detection

B. Take a batch of the same matrix material as for A. without spiking, with an amount sufficient for taking the planned number of samples
At least 8 laboratory samples with the minimum required amount of material are taken and analysed in duplicate: n; the number of results is at least 16
↳ analyse the presence or absence of the target in the samples: TN and FP

Standard statistics for binary results (yes/no, positive/negative, etc.) are correctness, sensitivity and specificity to be calculated form a set of results obtained from an array of samples. The correctness is the fraction of correct results, either positive or negative. This will be expressed as a correctness coefficient (*CS*) in order to avoid confusing with the general application of the term Accuracy. The frequently applied equation remains the same. The sensitivity (SE) is the ability of the method to detect the contaminant when it is present, whereas the specificity (*SP*) is the ability to not detect the contaminant when it is absent. In the case of sensitivity reasonable low levels should be tested, which should be below the legal limit. A matrix effect can influence the performance of the method, e.g. a compound feed with a high amount of minerals which would "dilute" bone fragments in the larger mass of sediment material. Specificity can be influenced by mimicking particles, such as salmon meal for terrestrial animal particles. Some of these factors should be included in additional samples, one at the time.

The following equations are designed to calculate the statistics:

$$CS = \frac{TP+TN}{TP+FN+FP+TN} \quad (18)$$

$$SE = \frac{TP}{TP+FN} \quad (19)$$

$$SP = \frac{TN}{FP+TN} \quad (20)$$

where TP is the number of correct positive identifications (true positives), TN the number of correct negative identifications (true negatives), FP the number of false positives and FN the number of false negatives. The statistics are presented as fractions or as percentage. Correctness (specificity or sensitivity) has to be calculated for each sample type independently.

A threshold of 0.95 or 95% for either sensitivity or specificity can be applied as criterion for a good or excellent score. However, one error in 16 results, the proposed minimum number of results, would give a score of 0.9375 or 93.75%. One error in 20 results would give a score of 0.95, at the limit of the proposed threshold. Statistic modelling indicated that 59 results would lead to a 95% lower confidence limit of a correct result exceeding the threshold of at least 19 correct out of 20 results.

3.2.2 Detection limit: sensitivity

As argued in Part 1: Theory and Principles, both decision limit CCα and detection limit CCβ cannot be calculated for qualitative methods in the classical sense (i.e. as for analytical chemical methods). An alternative procedure could be to analyse at least 20 samples at a low spike level as indicated in the box of paragraph 3.2.1. If the coefficient for sensitivity (equation 19) is 95% or above, the method can be declared fit for detection at that tested level. There is no prove that the method would be fit at lower levels, nor an indication that it should not be fit. Several levels can be tested simultaneously. Therefore, if any threshold is available (e.g. at least 0.1% of processed animal proteins should be successfully detected), any of the intended spike levels of a validation study as expressed in 3.2.1 should be at or below that threshold. This is likely to be organised in the form of an interlaboratory study, but such an approach needs sufficient approval of the required expertise level of the participants before the test samples are being distributed for validation.

3.2.3 Selectivity

A. Produce a homogenised batch of sample material with the target at the intended spike level, with an amount sufficient for taking the planned number of samples; add a mimicking material to the batch and homogenise
At least 8 (n) laboratory samples with the minimum required amount of material are taken and analysed in duplicate; the number of results is at least 16 ($2*n$)
↳ analyse the presence or absence of the target in the samples: PA and ND
An additional experiment can be performed with an alternative mimicking substance, if desired

B. Take a batch of the same matrix material as for A. without spiking, with an amount sufficient for taking the planned number of samples; add a mimicking material to the batch and homogenise
At least 8 laboratory samples with the minimum required amount of material are taken and analysed in duplicate: n; the number of results is at least 16
↳ analyse the presence or absence of the target in the samples: NA and PD

Where the basic results in terms of Correctness are primarily collected under repeatability circumstances including the absence of a mimicking material or ingredient, the same parameters can be calculated in the *presence* of such a look-alike material. Fish meal is the most commonly used ingredient in compound feed that could act as mimicking material for the detection of terrestrial animal proteins. Correctness, sensitivity and specificity (equations 18, 19 and 20) will be calculated independently for every specific type of substance, for example terrestrial animal material, fish material, insect material, based on the results of a series of samples per treatment.

3.2.4 Repeatability and reproducibility: accordance and concordance

Accordance is the probability of finding identical results in pairs of replicates of the same treatment in the same laboratory under repeatability conditions. This is equivalent to repeatability for quantitative results. Accordance is calculated with the equation:

$$ACC = \frac{\sum_1^n (p_{0,i}^2 + p_{1,i}^2)}{n} \quad (21)$$

with n is the number of samples with each a set of replicates, $p_{1,i}$ is the fraction of replicates with result 1 (boolean for presence) for sample i, and $p_{0,i}$ is the fraction of replicates with result 0 (boolean for absence) for sample i. The sum $p_{0,i} + p_{1,i}$ equals one, which implies that the choice for the coding of the fractions, either 0 or 1, is arbitrary.

There are several approaches for retrieving data for the calculation of Accordance and Concordance:
A. Take the data as collected in experiment A for the calculation of Correctness (section 3.2.1).
↳ analyse the presence or absence of the target in the pairs of duplicates: p_i

B. Produce a homogenised batch of sample material with the target at the intended spike level, with an amount sufficient for taking the planned number of samples
At least 8 (n) laboratory samples with the minimum required amount of material are taken and analysed in sets of three or more replicates: m; the number of results is at least 24 (m*n)
↳ analyse the presence or absence of the target in the pairs of duplicates: p_i

NOTE: the circumstances of the experiments determine the use of the data: repeatability circumstances allow to calculate Accordance, and reproducibility circumstances allow to calculate Concordance

Concordance is the probability of finding the same result for the same treatment under reproducibility circumstances. This situation can be achieved by applying a time interval or different technicians for the examination of the replicates. Concordance is calculated from the chance of finding pairs of replicates from any differentiating circumstance with identical results, either 1 (p) or 0 (1 – p):

$$CON = \bar{p}_1^2 + \left(1 - \bar{p}_1\right)^2 \quad (22)$$

where $\bar{p}$ is the average fraction of replicates with result "1", calculated as $(\sum_1^n p_{1,i})/n$.

The closer the value of both accordance and concordance to one the better the repeatability or reproducibility, respectively.

Explanation
The data for accordance and concordance need to be collected under repeatability or reproducibility circumstances, respectively. The following examples are intended to show the effect of one or more errors. These effect will be used to explain the behaviour of the two parameters.

Let us assume a dataset based on ten samples, each examined for two replicates. Table 5 shows six different scenario's with no, one, two or three errors.

Table 5 *Six scenarios for evaluating sets of 20 ordinal observations. Parameters used are Correctness (equation 14), Accordance (equation 17) and Concordance (equation 18).*

	A	B	C	D	E	F
number of errors	0	1	2	2	3	3
distribution	random	random	random	sample	random	mixed
Correctness	1.00	0.95	0.90	0.90	0.85	0.85
Accordance	1.00	0.95	0.90	1.00	0.85	0.95
Concordance	1.00	0.905	0.82	0.82	0.745	0.745
pairs of result	1 -1	1 -1	1 -1	1 -1	1 -1	1 -1
per sample	1 -1	1 -1	1 -1	1 -1	1 -1	1 -1
	1 -1	1 -0	1 -1	1 -1	0 -1	1 -1
	1 -1	1 -1	0 -1	0 -0	1 -1	1 -1
	1 -1	1 -1	1 -1	1 -1	0 -1	1 -1
	1 -1	1 -1	1 -1	1 -1	1 -1	0 -0
	1 -1	1 -1	1 -0	1 -1	1 -1	1 -1
	1 -1	1 -1	1 -1	1 -1	1 -0	1 -1
	1 -1	1 -1	1 -1	1 -1	1 -1	1 -0
	1 -1	1 -1	1 -1	1 -1	1 -1	1 -1

- Accordance, data is collected under repeatability circumstances: the value for Accordance equals the value for Correctness in all cases where the errors are randomly distributed over the samples (Table 5, columns B, C and E). A value for Accordance exceeding the value for Correctness indicates a situation where one or more pairs with false results exist (columns D and F). those pairs of repeated false results might indicate a systematic error.
- Concordance, data is collected under reproducibility circumstances: the value for Concordance will be lower with more errors. The lowest possible value is 0.5, indicating an equal share of correct or false results. With a number of false results exceeding a share of 50% the value for Concordance will increase, since false results are increasingly reproduced correctly.

3.2.5 Robustness/stability

Produce a homogenised batch of sample material with the target at the intended spike level, with an amount sufficient for taking the planned number of samples; this batch can be combined with the batch under action A of paragraph 3.2.1

At least 4 laboratory samples with the minimum required amount of material are taken and analysed in duplicate: k; the number of results is at least 8

↳ analyse the presence or absence of the target in the samples following the procedures of the method to be tested with a modification for one parameter: PA and ND

An additional experiment can be performed with an alternative modified parameter, if desired

The robustness of a method can be documented by investigating samples according to the fixed method, except for a modification of one factor. The usual statistics as applicable for quantitative methods cannot be applied to qualitative methods. Still factors exist which can influence the performance of a method. Documentation can be collected from experiments under modified circumstances. Since for qualitative methods frequency of correct results per sample is the principle result, at least 20 samples should be investigated per factor. The parameters correctness, sensitivity and specificity (equations 18, 19 and 20) can be used to calculate possible underperformance. If pairs of test portions taken from a sample can be used, the pairs of results can be analysed with the parameter accordance (equation 21) under the modified circumstance.

3.3 Identification

Proper identification is an issue for most visual methods, since an identity or classification is commonly established by means of the expertise of the technician. A conclusion of correct identity of a material should ideally be based on indications for Accuracy and Precision. These quality parameters can only be established when specific prerequisites are met. A group of methods which could be formally validated produce binary results: a visual response can be observed for the indication of an *a-priori* specified identity, and the response is not given for all other identities. Examples are the detection of blood particles using tetramethylbenzidine, Nile Red for plastics, and the hydroxide/ethanol based colour reaction for ergot sclerotia (Method IAG-A4). Alternatively, identification procedures of feed and food materials can result in any of a large array of identities. Examples are the identification of feed materials in compound feeds and the identification of weed seeds in bird feeds or in granular cereal samples. The validation of these two types of methods for establishing identity will be discussed in the following two sections.

3.3.1 Validation of binary identification methods

Validation of what is indicated here as binary identification methods requires the presence of a procedure producing an a-priori defined response. The procedure to follow for establishing the identity of a sample material relies on the documentation of that sample. The initial identity should not be known when the procedure for identification is performed, but it should preferably be available for comparison afterwards. If the sample is documented as reference material or is documented otherwise as having an approved identity, a replication of that identity is a confirmation of both Accuracy and Precision. If the formal identity is not known, either because the initial identity is not approved or is not available, then only precision can be established. The identification of ergot sclerotia by the hydroxide/ethanol based colour reaction is presented in Appendix 1 as example for this type of methods.

Three experimental designs can exist:

I. Formal identity is available, a new identification is intended to confirm this.
II. Initial identity is not approved, a new identification is intended to reproduce this.
III. An initial identity is absent, two independent identifications are intended to be conform.

An identity is based on a set of statuses, one for each characteristic. A status can be, e.g. hairs [present|not_present], lacunae [oval|not_oval], length [larger_then_# mm|not_larger_then_#_mm], colour [brown|not_brown], etc. For the current purpose, multistate characters (colour ranges, variable length) are reduced to binary (two-state) characters, i.e. the characteristics are treated as Booleans. Validation in terms of Accuracy or Precision can be achieved for the identity as a whole or for each of the included characteristics. Again, for the current purpose, a procedure for the validation of single characters will be followed.

Ordinal values: Cohen's Kappa

Each sample of a set of samples will be identified by using a formalised procedure for establishing the state of the chosen characteristic, either by expert judgment or by using an expert system. In each of the three described designs I., II. and III., two results per sample for the chosen characteristic are available. Then four possible combinations can occur:

a. The two known and/or observed statuses are equal presences.
b. The two known and/or observed statuses are not equal: not_present versus present.
c. The two known and/or observed statuses are not equal: present versus not_present.
d. The two known and/or observed statuses are equal not_presences.

The situations b) and c) represent two different combinations in design I. Situations a) and c) share the a-priori approved presence of a characteristic, and b) and d) share the a-priori approved absence (not_present) of that characteristic. There is no difference between situation b) and c) in design III.: both identifications have identical weights in the comparison and are achieved in random order. A logic table for the chosen characteristic will be filled consisting of the four observed frequencies, as shown in Table 6. In order to conclude a correct performance of the identification process, information should be collected on the presence of the chosen character for correct identification of the target as well on the absence of the chosen

character for correct identification of nontarget material. Numerically, the column sums *r* and *s* should both be higher than zero and should equal each other as close as possible. This situation would give a balanced situation for the evaluation of the numbers of deviating samples *b* and *c*.

Table 6 *Error matrix with frequencies of the combined output of two observations for a bistate character.*

Initial character state or first observation: Final observation:	present	not_present	 Row sums
present	*a*	*b*	*p=a+b*
not_present	*c*	*d*	*q=c+d*
Column sums	*r=a+c*	*s=b+d*	*n=p+q=r+s*

The figures as compiled in the logic table will be used to calculate two measures. The first one is based on the similarities or agreements *a* and *d*. In an ideal case the sum of *a* and *d* equals *n*. The parameter p_o will then equal one:

$$p_o = \frac{a+d}{n} \tag{23}$$

However, errors can occur, and above that, a part of the similarities will be agreements by chance. This stochastic influence will be expressed in the parameter p_e by using the dissimilarities *b* and *c*:

$$p_e = \left\{\left(\frac{a+b}{n} * \frac{a+c}{n}\right) + \left(\frac{c+d}{n} * \frac{b+d}{n}\right)\right\} = \left\{\frac{p*q}{n} + \frac{r*s}{n}\right\} \tag{24}$$

Finally, Cohen's Kappa statistic indicates the reproducibility of the chosen character status:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{25}$$

A conclusion on Precision (reproducibility) is the only result for situations II. and III. Considering the option that the first identification can be wrong, this wrong result can be confirmed by a second wrong identification. This combination of equally wrong identifications is not corrected by using the parameter p_e. As far as the a-priori available identity has been approved to be correct in situation I, both Accuracy (indicated as trueness) and Precision (reproducibility) are validated. A threshold for κ can be chosen at a level of 0.95.

The whole set of characters, together indicating a certain identity, can be chosen for validation. In principle a multi-class situation applies. For example, there are numerous types of feed ingredients and a range of different sources for processed animal proteins. The application of a two-class system with the two classes e.g. "soya_expeller" and "non_soya_expeller" is only realistic when the reproducibility of the identification "soya expeller" is aimed. A two-class system "ruminant" and "non_ruminant" for identification of bone fragments would make sense in a legal framework. There are statistics for testing multi class systems.

Concordance can be calculated in those cases where reproducibility circumstances apply. This could in particular be the case when reference material, identified in the past, has been used (design I.). Other options for verifying the quality of the identification is the calculation of Chi-square (X^2) or the Fisher Exact test.

Quantitative data: rank order and t tests

For quantitative results indicating an identity the Mann-Whitney U test (non-parametric, independent samples), Wilcoxon signed rank test (non-parametric, matched results) or paired or unpaired t-tests (parametric) can be applied.

3.3.2 General approach for establishment of identity and authenticity

Food and feed materials can have a large diversity of identities and a range of hand books are available for establishing the correct identity, which are listed in Part 1, section 5.2.1. Validation of a correct identification can follow two approaches.

Validation based on a subset of materials
Some materials are much more frequently applied than a large range of other materials. In compound feeds, by-products of plant oil production are predominantly originating from soya, rape seed, sunflower, palm or citrus. A technician can prove his or her competence by identifying a chosen subset of materials with a-priori agreed identity, which is not known to the technician (blind samples). This subset should consist of the materials with a high frequency among the samples from practice, and preferably in duplicate or triplicate. Thresholds for acceptable performance could be 90% or 95% correct identifications. Be aware that this is not a validation of the method but rather a validation of the expertise of the technician. Change in the procedure of material handling is less radical than change of technician. Every technician should prove his or her own competence, but once established, there is only a need to repeat this test after a long period of inactivity.

Validation by expert judgment
Another option to approve the correctness of an identity is to compare the judgment of two technicians, who independently examined the sample or material to be identified. This is not a validation procedure in strict sense, since it is part of the method, but can act as such in those situations that validation of a method prior to practical application is not possible. In a range of laboratory settings only one technician is present. An alternative in those cases is the application of expert systems: the technician enters a set of observations into the expert systems, and a list of most probable identities will be produced. A log of the observations and most probable identities, produced by the expert system, should be saved for later confirmation. Part 1 section 5.4.1 and Appendix 3 provide further information.

Expert judgment can also be applied for quality control at regular intervals after the validation of the competence of a technician. In the case of a one-technician laboratory setting, sets of already identified samples from practice can be sent to another laboratory at regular intervals for approval, for example biannual or quarterly.

4 Skewed data distribution

Data resulting from analysis might be skewed because they include principally non-negative numbers. Especially at lower contamination levels a short left tail (towards the value zero) or none at all might occur. In clearly skewed or suspicious situations a logarithmic transformation can sometimes be helpful. It can be applied if there are no values equal to zero. The individual data points will be transformed using a ^{10}log or natural log function (ln). After calculating the desired quality parameter, the reverse transformation (10-to-the-power or exponential function) will be carried out.

The following example is a calculation of the (mean) recovery R, based on the equation (1). An alternative would be to calculate the geometric mean recovery R_{GM}. Every value r_i and s_i is logarithmically transformed:

$$r_{ln} = \frac{\sum(\ln(r_i) - \ln(s_i))}{n} \tag{26}$$

The intermediate value r_{ln} can be transformed to an adjusted version of the usually applied parameter R by:

$$R_{GM} = 100 * e^{r_{ln}} \tag{27}$$

Appendix 1 Example cases

Three cases of validation of a specific visual inspection method have been translated to calculation sheets. These three examples will be presented and discussed briefly in the next section of this Appendix. The calculation sheets in Excel are available upon request.

Quantitative methods: *Ambrosia* validation
The method as presented in Part 1, Chapter 5.4.1 (examination of two portions of approx. 500 gram each, after a result above a threshold examination of a second set of 2 portions), is the situation matching Regulation (EU) No 691/2013, amending Regulation (EC) 152/2009 in terms of the amount of material to be examined. This procedure also allows to collect documentation for quality control. In the process of revision of Regulation (EC) 152/2009 a simplified version was agreed in 2021. In this version, one portion will be examined and in the situation of exceedance of a threshold a second portion will be examined.

These procedures are intended to collect documentation for quality *control*. Besides this, *validation* is aiming at approving the sampling handling supporting the examinations, and the technical skills for performing those examinations, as can be seen from the descriptions of the quality parameters. Therefore, the validation of the method to detect *Ambrosia* seeds in bird feed will be based on samples of 500 grams which are individually spiked and examined. The procedures as explained in the boxes for the respective parameters have been followed for collecting the relevant data. Values for the parameters Recovery, Quantification limit, Selectivity, Reproducibility and Robustness have been calculated according to the equations of section 3.1.

The next page shows an overview of the steps in the calculations for quality parameters as implemented in an Excel sheet. This Excel file is available, and contains a tab Overview, a tab Weight, a tab Counts and a tab Raw data.

Green cells: entered data
Grey cells: final values for parameters
Equation number refer to the equations in Part 2

Validation Ambrosia: seed weights												
3.1.1: Recovery 4 seeds/kg	A	B	C	D	E	F	G	H	average	n		
sample weight (g)	501	502	499	500	498	503	501	499	500.38	8		
spiked (mg)	7.6	9.3	9.8	8.9	8.6	10.1	8.5	8.9				
recovered (mg)	7.5	9.2	9.7	8.7	8.8	9.9	4.5	8.9				
recovered (mg/kg)	15.0	18.3	19.4	17.4	17.7	19.7	9.0	17.8				
T_c	98.7	98.9	99.0	97.8	102.3	98.0	52.9	100.0				
average spike	9.0											
average recovered	8.4											
R = sum (Tc) / n	93.45	%	**(Eq. 1)**	R								
SD Recovery	16.43	%	**(Eq. 2)**	SD_R								
3.1.2: Quantification limit												
$\lvert r_i - s_i \rvert$	0.1	0.1	0.1	0.2	0.2	0.2	4.0	0.0			based on data from experiment 3.1.1, low level	
$(r_i - s_i)^2$	0.010	0.010	0.010	0.040	0.040	0.040	16.000	0.000				
sum $(r_i - s_i)^2$	16.15											
$\delta = (\text{sum}\lvert r_i - s_i \rvert) / (k*w)$	1.22	mg/kg	**(Eq. 3)**	δ								
S_{DL}	1.52		**(Eq. 4)**									
$DL = 3 * S_{DL}$	9.11	mg/kg	**(Eq. 5)**	DL								
3.1.1: Recovery 12 seeds/kg	A	B	C	D	E	F	G	H	average	n		
sample weight (g)	501	498	499	503	499	501	498	502	500.13	8		
spiked (mg)	24.5	26.9	22.3	23.8	24.5	27.2	25.9	24.8				
recovered (mg)	24.5	26.3	22.1	23.9	24.5	18.3	25.8	24.9				
recovered (mg/kg)	48.9	52.8	44.3	47.5	49.1	36.5	51.8	49.6				
T_c	100.0	97.8	99.1	100.4	100.0	67.3	99.6	100.4				
average spike (mg)	25.0											
average recovered (mg)	23.79											
R = sum (Tc) / n	95.57	%	**(Eq. 1)**	R								
SD Recovery	11.47	%	**(Eq. 2)**	SD_R								
3.1.3: Selectivity	A	B	C	D	E	F	G	H	average	k		
sample weight (g)	500	498							499	2		
spiked (mg)	45.3	42										
recovered (mg)	45.1	42.5										
recovered (mg/kg)	90.2	85.3										
T_c	99.6	101.2										
average spike	43.7											
average recovered	43.8											
R = sum (Tc) / n	100.34	%	**(Eq. 6)**	R								
$\lvert r_{is} - s_{is} \rvert$	0.2	0.5										
sum $(\lvert r_{is} - s_{is} \rvert)/k$	0.35	mg										
$\delta = (\text{sum}\lvert r_{is} - s_{is} \rvert) / (k*w)$	0.70	mg/kg	**(Eq. 7)**	δ								
3.1.4: Reproducibility	A	B	C	D	E	F	G	H	average	n		
sample weight (g)	501	502	499	500	498	503	501	499	500.38	8	data from experiment 3.1.1, low level, first repetition	
	501	500	499	498	498	502	498	499	499.38		data from experiment 3.1.1, low level, second repetition	
d1	7.5	9.2	9.7	8.7	8.8	9.9	4.5	8.9				
d2	7.4	9.1	9.5	8.4	9.0	9.5	8.4	8.6				
average	7.450	9.150	9.600	8.550	8.900	9.700	6.450	8.750				
difference d	0.10	0.10	0.20	0.30	-0.20	0.40	-3.90	0.30				
difference d^2	0.01	0.01	0.04	0.09	0.04	0.16	15.21	0.09				
difference SQRT((d^2)/2)	0.07	0.07	0.14	0.21	0.14	0.28	2.76	0.21				
rsd-Ri (%)	0.95	0.77	1.47	2.48	1.59	2.92	42.76	2.42				
average overall	8.57											
sum d^2	15.65											
SR_w= SQRT(SUM d^2 / 2k)	0.99		**(Eq. 10)**	SRw								
3.1.6 Robustness	A	B	C	D	E	F	G	H	average	k		
sample weight (g)	501	499							500	2		
spiked (mg)	45.3	48.9										
recovered (mg)	43	47										
$\lvert r_{is} - s_{is} \rvert$	2.3	1.9										
sum $(\lvert r_{is} - s_{is} \rvert)/k$	2.1	mg										
$\delta = (\text{sum}\lvert r_{is} - s_{is} \rvert) / (k*w)$	4.2	mg/kg	**(Eq. 13)**	δ								
$\Delta\delta s = \delta^* s - \delta$	2.98	mg/kg	**(Eq. 14)**	$\Delta\delta s$								

Validation Ambrosia: seed number

3.1.1: Recovery 4 seeds/kg	A	B	C	D	E	F	G	H	average	n	
sample weight (g)	501	502	499	500	498	503	501	499	500.375	8	
spiked (sd)	2	2	2	2	2	2	2	2			
recovered (sd)	2	2	2	2	2	2	1	2			
recovered (sd/kg)	4	4	4	4	4	4	2	4			
T_c	100.0	100.0	100.0	100.0	100.0	100.0	50.0	100.0			
average spike (sd)	2										
average recovered (sd)	1.875										
R = sum (Tc) / n	93.75 %	**(Eq. 1)**		R							
SD Recovery	17.68 %	**(Eq. 2)**		SDR							
3.1.2: Quantification limit											
$\|r_i - s_i\|$	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0			based on data from experiment 3.1.1, low level
(yi - mi)^2	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000			
sum (yi - mi)^2	1.00										
$\delta = (\text{sum}\|r_i - s_i\|) / (k*w)$	0.25 sd/kg	**(Eq. 3)**		δ							
S_{DL}	0.38	**(Eq. 4)**									
$DL = 3 * S_{DL}$	0.00 sd/kg	**(Eq. 5)**		DL							

3.1.1: Recovery 12 seeds/kg	A	B	C	D	E	F	G	H	average	n	
sample weight (g)	501	498	499	503	499	501	498	502	500.125	8	
spiked (sd)	6	6	6	6	6	6	6	6			
recovered (sd)	6	6	6	6	6	5	6	6			
recovered (sd/kg)	12.0	12.0	12.0	11.9	12.0	10.0	12.0	12.0			
T_c	100.0	100.0	100.0	100.0	100.0	83.3	100.0	100.0			
average spike (sd)	6										
average recovered (sd)	5.875										
R = sum (Tc) / n	97.92 %	**(Eq. 1)**		R							
SD Recovery	5.89 %	**(Eq. 2)**		SDR							

3.1.3: Selectivity	A	B	C	D	E	F	G	H	average	k	
sample weight (g)	500	498							499	2	
spiked (mg)	10	10									
recovered (mg)	10	9									
recovered (mg/kg)	20.0	18.1									
T_c	100.0	90.0									
average spike	10.0										
average recovered	9.5										
R = sum (Tc) / n	95.00 %	**(Eq. 6)**		R							
$\|r_{is} - s_{is}\|$	0	1									
sum ($\|r_{is} - s_{is}\|$)/k	0.5 n										
$\delta = (\text{sum}\|r_{is} - s_{is}\|) / (k*w)$	1.00 sd/kg	**(Eq. 7)**		δ							

3.1.4: Reproducibility	A	B	C	D	E	F	G	H	average	n	
sample weight (g)	501	502	499	500	498	503	501	499	500.375	8	data from experiment 3.1.1, first repetition
	501	498	499	503	499	501	498	502			data from experiment 3.1.1, second repetition
d1 (sd)	2	2	2	2	2	2	1	2			
d2 (sd)	2	2	2	2	2	2	2	2			
average	2.000	2.000	2.000	2.000	2.000	2.000	1.500	2.000			
difference d	0.00	0.00	0.00	0.00	0.00	0.00	-1.00	0.00			
difference d^2	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00			
difference SQRT((d^2)/2)	0.00	0.00	0.00	0.00	0.00	0.00	0.71	0.00			
rsd-Ri (%)	0.00	0.00	0.00	0.00	0.00	0.00	47.14	0.00			
average overall	1.94										
sum d^2	1.00										
SR_w= SQRT(SUM d^2 / 2k)	0.25	**(Eq. 10)**		SRw							

3.1.6 Robustness									average	k	
	501	499							500	2	
spiked (sd)	10	10									
recovered (sd)	9	10									
$\|r_{is} - s_{is}\|$	1	0									
sum ($\|r_{is} - s_{is}\|$)/k	0.5 n										
$\delta = (\text{sum}\|r_{is} - s_{is}\|) / (k*w)$	1.00 sd/kg	**(Eq. 13)**		δ							
$\Delta\delta s = \delta*s - \delta$	0.75 sd/kg	**(Eq. 14)**		Δδs							

Qualitative methods: animal proteins validation

A frequently used qualitative method is the monitoring for animal proteins in feed or in feed ingredients (Regulation (EC) 152/2009, Annex VI, Part I, and connected SOPs). This method is designed to detect simultaneously particles of terrestrial animal origin and of fish. The extension with double sedimentation for the detection of insects parts is available in draft and is assumed to be in force late 2021 or in 2022.

Usually proficiency tests are intended to collect documentation on both targets of the current implementation. Tables in PT reports show values for Correctness, Sensitivity and Specificity separated for the two targets terrestrial animal and fish material, and in some cases combined or overall values for the three performance parameters. A validation study is intended to document one target, as indicated in the experimental boxes. Terrestrial animal material is the target of the following example of a validation study of qualitative methods, with fish as mimicking or confusing material.

Five experiments with eight samples each have been carried out in this example of a validation study. For each experiment a batch of homogenised material have been produced, and eight samples have been taken from each of these five batches with the amount of material as desired in the method (50 grams). Every sample is analysed twice in a short time frame; these replicates are NOT the two determination cycles which are sometimes necessary according to the method. The total number of analyses is 80 (5 x 8 x 2).

The experiments are:

1. Eight analyses in duplicate of a blank matrix (paragraph 3.2.1 B)
2. Eight analyses in duplicate of a matrix spiked with terrestrial animal material at low level (0.02%; paragraph 3.2.1 A)
3. Eight analyses in duplicate of a matrix spiked with terrestrial animal material at high level (0.1%; paragraph 3.2.1 A)
4. Eight analyses in duplicate of a matrix without the target and spiked with fish meal (2%; paragraph 3.2.3 B)
5. Eight analyses in duplicate of a matrix spiked with both terrestrial animal material and with fish meal (0.1% and 2%, respectively; paragraph 3.2.3 A)

Additionally, the results of experiment 2) have been used to calculate the Accordance based on the assumption that the replicates are analysed under repeatability conditions (approach A of paragraph 3.2.4). If the performance of experiment 2) in terms of sensitivity is exceeding the a-priori fixed level, the detection limit can be declared to be at least 0.02% (paragraph 3.2.2). In this case with a value of SE = 0.88 the method can be assumed to show underperformance for sensitivity at the low level.

The experiments 1), 3), 4) and 5) include every possible combination of terrestrial animal material and fish material spiked at 0.1% and 2%, respectively. This means that the same results can be used for calculating the performance for fish meal, rearranged with fish meal as target material and terrestrial animal material as mimicking material. A copy of the same sheet can be used for the calculation. In this way only one spike level is available for the spike material (2%). If desired, experiment 2) (low level of target material, now fish meal), can be performed additionally in order collect full data.

Green cells: entered data (1: correct result; 0: incorrect result)
Grey cells: final values for parameters
Equation number refer to the equations in Part 2

Validation animal proteins, addition fish meal						correct results expressed as 1				
3.2.1: blank	A	B	C	D	E	F	G	H	n	blank
result 1, target	1	1	1	1	1	1	1	1	8	
result 2, target	1	1	1	1	1	1	1	1		
Positive deviations	0									
Negative agreements	16									
Specificity	1.00	**(Eq. 17)**								
3.2.1: low level	A	B	C	D	E	F	G	H	n	target low level, no mimicking material
result 1, target	1	1	1	1	0	1	1	1	8	
result 2, target	1	1	1	1	0	1	1	1		
Positive agreements	14									
Negative deviations	2									
Sensitivity	0.88	**(Eq. 16)**								
Correctness	0.94	**(Eq. 15)**								
3.2.1: high level	A	B	C	D	E	F	G	H	n	target high level, no mimicking material
result 1, target	1	1	1	1	1	1	1	1	8	
result 2, target	1	1	1	1	1	1	1	1		
Positive agreements	16									
Negative deviations	0									
Sensitivity	1.00	**(Eq. 16)**								
Correctness	1.00	**(Eq. 15)**								
3.2.3: without target	A	B	C	D	E	F	G	H	n	no target, mimicking material
result 1, target	1	1	1	1	1	1	1	1	8	
result 2, target	1	1	1	1	1	1	1	1		
Positive deviations	0									
Negative agreements	16									
Specificity	1.00	**(Eq. 17)**								
3.2.3: with target	A	B	C	D	E	F	G	H	n	target, mimicking material
result 1, target	1	1	1	1	1	0	1	1	8	
result 2, target	1	1	1	1	1	1	1	1		
Positive agreements, target	15									
Negative deviations, target	1									
Sensitivity	0.94	**(Eq. 16)**								
Correctness	0.97	**(Eq. 15)**								
3.2.4: Accordance, low level	A	B	C	D	E	F	G	H	n	target low level, no mimicking material
result 1, target	1	1	1	1	0	1	1	1	8	data from experiment 3.2.1, low level
result 2, target	1	1	1	1	0	1	1	1		
n_i	2	2	2	2	2	2	2	2		
n_{1i}	2	2	2	2	0	2	2	2		
n_{0i}	0	0	0	0	2	0	0	0		
p_{1i}	1	1	1	1	0	1	1	1		
p_{0i}	0	0	0	0	1	0	0	0		
p_{1i}^2	1	1	1	1	0	1	1	1		
p_{0i}^2	0	0	0	0	1	0	0	0		
$p_{1i}^2 + p_{0i}^2$	1	1	1	1	1	1	1	1		
$\Sigma (p_{1i}^2 + p_{0i}^2)$	8									
Accordance	1.00	**(Eq. 18)**								

Identification methods: colour reaction for ergot sclerotia identification
Method IAG-A4 for the Determination of Ergot (*Claviceps purpurea* Tul.) in Animal Feedingstuff presents a colour reaction for the confirmation of a correct identification of Ergot slcerotia. RIKILT has conducted in 2015 an experiment with 170 ergot sclerotia collected from two highly contaminated samples of rye. Six rye grains have been used as control. A subset of 42 ergot sclerotia and two rye grains have been tested for their alkaloid content (unpublished results).

The results have been entered in the performance sheet in two ways:
- A list of results of individual sclerotia. The calculation sheet will transfer these data to agreements (TRUE) or disagreements (FALSE) and to the values for Table 6 (cells a, b, c and d) of section 3.3.1. This way of entering was followed for the 42 sclerotia with alkaloid data and the six blanks.
- The data for Table 6 is entered as such. This way of entering was followed for the entire set of 170 sclerotia and six blanks.

The results show that the colour reaction for ergot sclerotia is not a reliable conformation of a visual inspection. A total of 23 particles, visually identified as ergot sclerotia, have been indicated as pseudo sclerotia (13.5%). Fourteen out of these 23 pseudo sclerotia did contain alkaloids. A set of 12 particles has been identified as sclerotia but appeared to contain no alkaloids or traces of it. This phenomenon has been documented (Mulder et al., 2012). In contrast, all sclerotia containing alkaloids have been identified as sclerotia after visual inspection (no false negatives).

The value for Kappa (section 3.3.1) has been calculated for both datasets. The dataset containing 42 sclerotia with indication of levels of alkaloids show a lower value for Kappa than the value for the entire data set containing 170 sclerotia (0.185 and 0.304, respectively). A more balanced number of sclerotia and of rye grains would have resulted in higher values for the Kappa statistic.

Additionally, the concordance was calculated (section 3.2.4). This statistic is exclusively based on mismatches (FALSE in the table) between the results of the visual identification and of the colour test. This statistic indicates an underperformance of the colour test.

Green cells: entered data (1: correct result; 0: incorrect result)
Grey cells: final values for parameters
Equation number refer to the equations in Part 2

Identity validation									
	Ergot sclerotia with alkaloid content					**Ergot sclerotia total**			
SAMPLE	visual	colour test	equal	combi	chemical (EA; µg/g)				
1	yes	yes	TRUE	a	7670				
2	yes	yes	TRUE	a	3527				
3	yes	no	FALSE	c	2766				
4	yes	no	FALSE	c	1818				
5	yes	no	FALSE	c	<5				
6	yes	no	FALSE	c	5898				
7	yes	yes	TRUE	a	4639				
8	yes	yes	TRUE	a	3706				
9	yes	no	FALSE	c	2973				
10	yes	no	FALSE	c	<2				
11	yes	no	FALSE	c	1901				
12	yes	yes	TRUE	a	1442				
13	yes	yes	TRUE	a	1865				
14	yes	no	FALSE	c	<2				
15	yes	yes	TRUE	a	4085				
16	yes	no	FALSE	c	1440				
17	yes	no	FALSE	c	1505				
18	yes	yes	TRUE	a	847				
19	yes	no	FALSE	c	2571				
20	yes	yes	TRUE	a	3871				
21	yes	yes	TRUE	a	1678				
22	yes	no	FALSE	c	<2				
23	yes	yes	TRUE	a	4833				
24	yes	yes	TRUE	a	<2				
25	yes	yes	TRUE	a	<2				
26	yes	yes	TRUE	a	240				
27	yes	yes	TRUE	a	<2				
28	yes	yes	TRUE	a	1230				
29	yes	no	FALSE	c	<2				
30	yes	no	FALSE	c	4292				
31	yes	yes	TRUE	a	428				
32	yes	no	FALSE	c	2372				
33	yes	no	FALSE	c	285				
34	yes	no	FALSE	c	<2				
35	yes	no	FALSE	c	2031				
36	yes	no	FALSE	c	<2				
37	yes	yes	TRUE	a	3085				
38	yes	yes	TRUE	a	<2				
39	yes	no	FALSE	c	2104				
40	yes	yes	TRUE	a	3773				
41	yes	no	FALSE	c	913				
42	yes	no	FALSE	c	<2				
43	no	no	TRUE	d	<2				
44	no	no	TRUE	d	<2				
45	no	no	TRUE	d	-				
46	no	no	TRUE	d	-				
47	no	no	TRUE	d	-				
48	no	no	TRUE	d	-				
basis			N	48			N	176	
			TRUE	26			TRUE	153	
3.3.1: Cohen's Kappa									
logic table			a	20			a	147	
			b	0			b	0	
			c	22			c	23	
			d	6			d	6	
probability of chance agreement		(Eq. 21)	Pe	0.438			Pe	0.812	
relative observed agreement		(Eq. 20)	Po	0.542			Po	0.869	
Kappa statistic *K*		(Eq. 22)	Kappa	0.185			Kappa	0.304	
			Pe_{yes}	0.365			Pe_{yes}	0.807	
			Pe_{no}	0.073			Pe_{no}	0.006	
			pe	0.438			pe	0.812	
			Kappa	0.185			Kappa	0.304	
column/row sums			n1,1	20			n1,1	147	
			n1,2	28			n1,2	29	
			n2,1	42			n2,1	170	
			n2,2	6			n2,2	6	
3.2.4: Concordance									
match (TRUE)				0.54				0.87	
nonmatch (FALSE)				0.46				0.13	
CON		(Eq. 19)		0.50				0.77	

Wageningen Food Safety Research
P.O. Box 230
6700 AE Wageningen
The Netherlands
T +31 (0)317 48 02 56
wur.eu/food-safety-research

WFSR Report 2022.006

The mission of Wageningen University & Research is "To explore the potential of nature to improve the quality of life". Under the banner Wageningen University & Research, Wageningen University and the specialised research institutes of the Wageningen Research Foundation have joined forces in contributing to finding solutions to important questions in the domain of healthy food and living environment. With its roughly 30 branches, 7,200 employees (6,400 fte) and 13,200 students and over 150,000 participants to WUR's Life Long Learning, Wageningen University & Research is one of the leading organisations in its domain. The unique Wageningen approach lies in its integrated approach to issues and the collaboration between different disciplines.

To explore the potential of nature to improve the quality of life

Wageningen Food Safety Research
P.O. Box 230
6700 AE Wageningen
The Netherlands
T +31 (0) 317 48 02 56
wur.eu/food-safety-research

WFSR report 2022.006

The mission of Wageningen University & Research is "To explore the potential of nature to improve the quality of life". Under the banner Wageningen University & Research, Wageningen University and the specialised research institutes of the Wageningen Research Foundation have joined forces in contributing to finding solutions to important questions in the domain of healthy food and living environment. With its roughly 30 branches, 7,200 employees (6,400 fte) and 13,200 students and over 150,000 participants to WUR's Life Long Learning, Wageningen University & Research is one of the leading organisations in its domain. The unique Wageningen approach lies in its integrated approach to issues and the collaboration between different disciplines.