Feasibility, Practicality and Usefulness of Sampling Proficiency Tests in the Food Sector (E01070)

Final report to FSA
December 2011

Project Investigators and authors

- Professor Michael. H. Ramsey, Principal Investigator
- Dr. Bas Geelhoed, Research Fellow

University of Sussex

Acknowledgements

The project investigators would like to thank the following people for their contributions to the project:

Dr. Andrew Damant	Project Officer, Food Standards Agency, London
Dr. Ellis Evans	Project Officer, Food Standards Agency, London
Dr. Roger Wood	Project Officer, Food Standards Agency, London
Prof. Mike Thompson	Birkbeck College, London
Debbie Dixon	An apple juice manufacturer, East Anglia
Roy Smyth	Rural Payment Agency
Martyn Allcorn	Westbury Creamery

Thanks are also extended to all the samplers and analysts who contributed to this project.

Contents

1. Summary

Two experimental Sampling Proficiency Tests (SPTs) have been held to demonstrate their feasibility, practicality and usefulness in the food sector. The first ever SPT on food looked at the determination of the toxin patulin in cloudy apple juice, using 9 trained participants to sample one batch of 6500 litres of unstirred juice from a tap on the side of the tank. It found that the expanded uncertainty on the measurements (19.4%) was so dominated by the precision of the analytical method, and the juice was so homogeneous, that it was impossible to detect differences between the performance of the samplers. The process of implementing this prototype test was used to identify 12 generic criteria that can be applied to design an effective SPT. These criteria were then used to design a second improved SPT, on the determination of moisture in fresh butter, which thereby overcame many of the limitations found in the first SPT. The precision of the analytical method was small enough to quantify the heterogeneity of the butter, and the proficiency of the samplers was reflected in their performance z-scores. Using a newly-devised scoring system, two samplers were found to be non-proficient. One of these non-proficient samplers was untrained and intentionally placed in the SPT to test its usefulness, but the other was a trained sampler. The results of the second SPT demonstrated not only the feasibility and practicality of SPTs, but also their usefulness in identifying non-proficient samplers. SPTs therefore have the potential to be used to improve the performance of samplers, in an analogous way to the way in which analytical proficiency tests (APTs) have already improved the performance of analytical laboratories over recent decades.

A further use of SPT results is to improve estimates of measurement uncertainty. It has already become accepted that the sampling procedure contributes to the uncertainty of measurement results in most circumstances, but previous methods of estimation did not include the contribution from sampling bias. The results of the second SPT showed that the measurement uncertainty estimated by the equivalent of the accepted ‘duplicate method’ (0.39%) was a factor of two lower than the more realistic value estimated by the SPT (0.87%), which does include substantial between-sampler bias. All of these findings have now been published in The Analyst[1].

2. Introduction

2.1 Relevance of this research to the food sector

The reliability of measurements of the chemical composition of food materials has become an important issue for regulators of food. Probably the most important quality parameter that estimates this reliability is the uncertainty of the measurements. It is now becoming recognized that this uncertainty arises from two main activities, the primary sampling of the food, and the chemical analysis of the resultant samples. Analytical proficiency tests (APTs) have already become a very valuable part of assuring appropriate quality in the chemical analyses made in analytical laboratories[2]. The logical extension of this approach is to also apply proficiency testing to the sampling part of the measurement process. This project aims to do this, for the first time for the food sector, using sampling proficiency tests (SPTs).

APTs consist essentially of the distribution and chemical analysis of a sufficiently homogeneous test material to all of the participants. Each participant then chemically analyses the sample for the specified analytes, by whatever routine analytical procedures they wish, and reports the results to the organizers. The organizers then calculate a performance score for each participant for each analyte, based upon how closely the reported measurement result agrees with a consensus value, judged in terms of a target standard deviation (or fitness-for-purpose) criterion. The results of the APT are useful, and probably essential, to participants in providing external assessment of the analytical quality control procedures within their own laboratory. In addition, APT results can also be used to make improved estimates of the uncertainty of the measurements, by including components that are often unsuspected, such as inter-organizational bias[3].

Sampling (including the physical preparation of samples), has now become accepted as an integral part of the measurement process, for most purposes[4]. The concept of a sampling proficiency test (SPT) was proposed in 1994 [5,6], and has since been successfully applied to the environmental sector to test the proficiency of samplers[7,8,9]. SPT results have also been used to estimate total measurement uncertainty in environmental studies[10,11], including the contributions from both the sampling and the sample preparation.

The logical progression is to consider whether the concept of the SPT is feasible in the food sector, and also whether it could be useful for improving the quality of food sampling. This requires the adaptation of the SPT methodologies to the special properties of food, such as short periods of stability. It also requires organizations and samplers in this sector to be convinced of the advantages of participating in an SPT. Having designed a food SPT, there are also logistical considerations to be overcome in locating a suitable sampling target (e.g. a batch of food) and gaining the agreement of the owner of this target to allow it to be used for this purpose. There is also a requirement for the further development of a suitable procedure for the scoring of participants in an SPT. This development is required because the current scoring method for APTs does not address the extra complications in an SPT caused by (i) the heterogeneity of the sampling target and (ii) the analytical uncertainty that might affect the measurements reported by the participants, and therefore potentially affect their SPT scores, even though it is not caused by their sampling proficiency.

Sampling proficiency testing is just one form of inter-organisational sampling trial (IOST)[9], but another is the collaborative trial in sampling (CTS)[6,12] , which has previously been applied once in the food sector. A CTS is aimed at the validation of a sampling protocol rather than assessing the proficiency of the sampler. In a CTS the participants are all

allocated the same protocol, rather than given the choice of which protocol to select as part of the test of proficiency, as in an SPT.

The one previous IOST reported in the food sector was a FSA-funded CTS applied to the determination of N, Mo and Pb in wheat, and moisture, caffeine and thiobromine in green coffee[13]. It successfully showed the feasibility of the concept of a CTS in the food sector, using large batches of wheat and green coffee as the sampling targets. It had five participants, including one professional sampler and four volunteers, who each received a brief training in the use of the sampling spear. This CTS did detect significant between-sampler variation for two analytes in the wheat (N and Mo). However, it lacked the statistical power to make further deductions, due to the low number of samplers. This shortcoming could be overcome by increasing the number of participants (in a future CTS or SPT) to at least the minimum number of eight[14].

There is a further potential benefit of food-based SPTs in that the results can be used for making more realistic estimates of the uncertainty of measurements, including the contribution from sampling bias. Current methods for the estimation of the uncertainty of measurements arising from sampling[15] have been shown to be effective in the food sector[16], but none of the methods applied so far can include the contribution from often unsuspected bias arising from the sampling process. The SPT has the potential to provide improved estimates of uncertainty that include this between-sampler bias, by incorporating the extra contribution from the between-sampler variance estimated from the SPT. As in APTs, a crucial first step is to identify the criteria that need to be met to make a test material (in this case a sampling target) suitable for use in an SPT.

In the broader context of assuring the appropriate quality of measurements, SPTs may also have a future role in the accreditation of organizations for sampling particular materials, or the certification of individual samplers. This is equivalent to the role that APTs already have in providing external assessment of a laboratory's proficiency for the purpose of accreditation (to ISO 17025). This may also provide a suitable motivation for more organizations to take part in SPTs, which will further enhance their usefulness.

2.2 Research objectives

The aim of this research is to study the feasibility, practicality and usefulness of Sampling Proficiency Tests in the food sector. Four scientific objectives were formulated to address this:

1) To devise criteria that can be used to select an effective target for a sampling proficiency test (SPT).
2) To apply these criteria to select two sampling targets for the design and implementation of two SPTs, in order to assess the general feasibility, practicality and usefulness of SPTs in the food sector.
3) To devise and apply a scoring scheme for SPTs, to give participants feedback on their performance
4) To use the SPT results to estimate uncertainty of measurement, and to compare these against estimates made using methods based upon single participants (e.g. the duplicate method).

2.3 Report aims and structure

A major deliverable of this research project is a peer-reviewed paper reporting on all four of these objectives, using the second case study on the determination of moisture in fresh butter[1] (attached as Appendix G).

The published paper has already described several aspects of the research that will consequently not be discussed in such depth in this report. These aspects include descriptions of the:

1. Previous development of SPTs and their potential for the evaluation of uncertainty,

2. Theory of scoring of SPTs, and evaluation of measurement uncertainty using SPT results,

3. Second case study and SPT, including its conformity to the general selection criteria, experimental design, results and discussion for scoring, and for uncertainty evaluation.

This report aims, therefore, to supplement this paper, by providing an overview of the relevance of this research to the food sector, and report on an initial first case study and SPT that was based upon the determination of patulin in cloudy apple juice. This initial SPT helped inform the design of the second study and also provided further evidence of the performance of SPTs. The report then seeks to discuss in more detail than the paper, the implications of all the research findings for the potential use of SPTs in the food sector in general. It also includes all of the raw data from both SPTs in Appendices (B-F, H-M) for future reference.

3. General selection criteria and design of SPTs

3.1 Criteria for design of an effective SPT

The criteria for the design of an effective SPT (Table 1 and in the paper in Appendix G) have been derived from consideration of the findings in the previous SPTs and from statistical considerations. The degree to which these selection criteria can be met will vary between different types of targets and in different sectors of application. Not every criterion has necessarily to be met in full, so the degree to which they need to be met will be discussed.

Table 1. Criteria for an effective SPT.

#	**Criteria**
1	Size of the target must be large enough to allow sampling by at least 8 samplers, without significantly affecting the target.
2	Presence of an analyte that can be determined with low enough analytical uncertainty to quantify sampling uncertainty.
3	Sufficient temporal stability of the sampling target (or else special procedures adopted to compensate for this)
4	A level of heterogeneity of the analyte in the sampling target that is detectable with the proposed method of chemical analysis
5	The presence of a recommended sampling protocol applicable to the selected analyte
6	Availability of appropriate samplers with the required technical skills
7	Potential usefulness for improving performance of the samplers, and accessibility of the sampling target
8	Required time to perform the sampling protocol and duration of the SPT
9	Acceptable level of cost due to damage to the target material and its packaging, during the SPT
10	Independence of the samplers participating in the SPT
11	Agreement of the owner of the sampling target and any sponsor of the SPT
12	Independence of the different contributors to the SPT

One example of the use of statistical considerations is the requirement that there be at least 8 samplers in an SPT (criterion 1), which is based upon the consideration of achieving a sufficiently small confidence interval on the variance estimates[14]. Temporal stability of the sampling target is very important (criterion 3), especially where spatial variability is being characterised[11]. However some systems, such as landfill gas, are naturally temporally variable and the design of the SPT can be modified to make allowance for this criteria not being met[9]. The ideal situation is still however that the composition of the target is kept constant over the duration of the SPT.

In most cases, a perfectly homogeneous sampling target will not allow an SPT to show differences between the performance of the samplers. One exception is bias caused by variable contamination of the samples, but bias from most other causes will not be detected, so sufficient heterogeneity is required to cover all possibilities. A minimum amount of heterogeneity of the analyte distribution in the target (criterion 4) is needed in a sampling target, therefore, for the different behaviour of the samplers to be quantified. The following study will demonstrate that the absolute value of the heterogeneity can be relatively low (e.g. 0.08% RSD), especially if the analytical method is very precise (e.g. 0.17 % RSD). This reflects therefore related requirement for the presence of an analyte that can be determined in the sample (criterion 2), with an acceptable detection limit (i.e. with an analytical uncertainty that enables the quantification of the between-sampler uncertainty).

The presence of a recommended sampling protocol applicable to the selected analyte (criterion 5), enables samplers to follow pre-designed written instructions. However, cases have been known where only verbal protocols were available, and the SPT was then able to investigate the effect of a lack of written instructions. The availability of appropriate samplers (criterion 6), refers not just to their number (at least 8, as in criterion 1), but also to their level of technical skill and their ability to participate at the required time and place. It is desirable that the samplers be properly trained, but assessing the effectiveness of any

training, especially if it is of short duration, can be a suitable objective for an SPT. A similar factor is to aspire to independence between the samplers (criterion 10). This applies during the SPT, when care needs to be taken so that no sampler can observe how another sampler is taking their samples. It also applies to the previous training, where ideally they should have not have been recently trained by the same person. The effects of joint training of the samplers could be investigated using an SPT, but the resultant estimates of uncertainty from this sampling protocol may be expected to be lower than that when independent samplers are operating.

The potential usefulness for improving performance of the samplers (criterion 7), refers to the degree of flexibility in the system. For example, all of the human samplers might use one mechanical sampling device, which may be badly designed and located. In this case the measurements of all of the samplers in an SPT may agree well, but they may all be unrepresentative of the bulk composition of the sampling target. Another example would be where the target material is kept in multiple sacks in a large pile, but the samplers can only access the sacks on the outside of the pile. This is an example of a poor sampling protocol that will affect the results of routine sampling as well as the SPT. The situation needs to be commented upon by the SPT organisers, and a different target chosen.

The time required to perform the sampling protocol (criterion 8) needs to not be excessive, so that the material does not change significantly in composition (criterion 3), and the cost to the SPT does not be thereby become unacceptable. This is closely related to the duration of the SPT, but obviously also depends on the number of samplers. There is also usually some level of damage to the target material, or its packaging, during the SPT (criterion 9). The SPT can be designed to reduce such damage as far as possible, whilst not deviating appreciably from the routine sampling protocol. The costs arising from the loss of target material during the SPT can be thereby minimized and included in the cost of running the SPT.

Clearly it is essential to secure the agreement of the owner of the sampling target, and any sponsor of the SPT (criterion 11). Practical experience has shown that the independence of the different contributors to the SPT (criterion 12) is also essential. For example, if the organisation that performs the chemical analysis is also a participant in the SPT, or the manufacturer or owner of the sampling target, then there may be a conflict of interest in reporting measurement that affect the commercial interest of that organisation. It is desirable, therefore, that all of these roles are kept in different organisations.

3.2 General experimental design of SPTs

The general experimental design used for most SPTs is a balanced design in which each sampler takes two samples, by independent interpretation of the protocol, both of which is analysed twice for each target analyte (Fig 1). The analytical duplicates enable a separation of the contribution to the uncertainty from the analytical repeatability. This is required to ensure that repeatability is small enough to enable the quantification of the sampling variances. The duplicated sampling by each sampler, provides an estimate of the within-sampler variance. The within-sampler variance is affected by the heterogeneity of analyte concentration within the sampling target. This heterogeneity is a key difference between and an SPT and an APT because in an APT heterogeneity is minimised as far as possible, but in an SPT it is an essential requirement (criterion 4 in Table 1). The experimental design has therefore to enable this heterogeneity to be quantified and separated so as not to affect the samplers' scores.

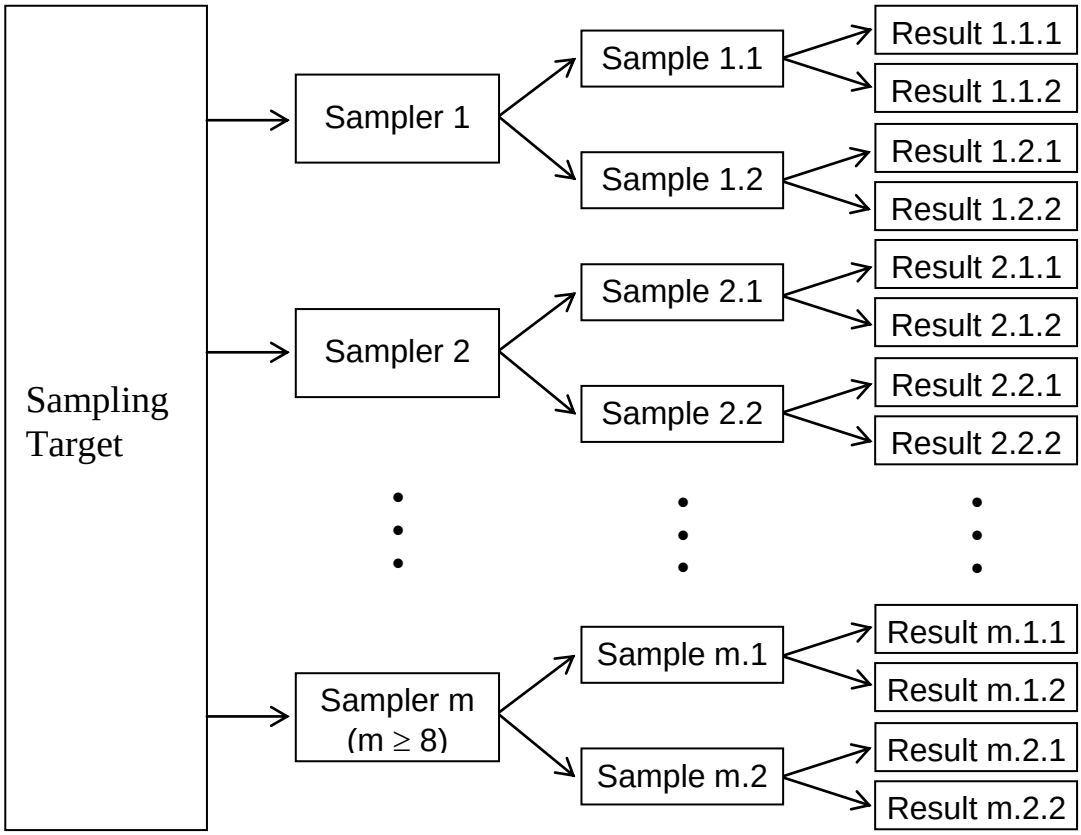

Figure 1. General experimental design applied sampling proficiency tests (SPT), and both SPTs in this research[1].

The between-sampler variance reflects the extra factors such as the between-sampler bias that can arise from systematic effects in either the sampling or the chemical analysis, depending on how the SPT is designed. The chemical analysis can either be undertaken by each of the participants separately, together with a matched reference material[7], or the samples from all of the participants can be collected centrally and chemically analysed in one laboratory under repeatability conditions[11]. The former approach could be called more precisely a 'measurement proficiency test' (MPT), as each participant undertakes the whole of the measurement process. The later approach eliminates the effect of any analytical bias between the participants, helping the performance score to reflect only the participant's sampling performance. However, the centralised analysis in the later approach will also tend to reduce the estimate of the total measurement uncertainty to an unrealistic level, unless it is dominated by the sampling component, or the analytical component is added independently.

4. First SPT on the determination of patulin in apple juice

4.1 Context for the SPT on patulin in apple juice

Mycotoxins occur naturally, being produced by certain fungi, and can cause a range of adverse health effects[17] that are of concern to the food sector. Patulin is one such mycotoxin, with a relatively simple chemical structure (Fig 2) and is produced by certain fungal species of *Penicillium, Aspergillus* and *Byssochlamps*[18]. Birkinshaw[19] first discovered Patulin in 1943 and initially believed it to have antibacterial properties, but it was later discovered to have toxic side effects. Patulin occurs naturally in various fruits including pears, grapes and bilberries, and traces have also been found in certain vegetables,

cereal grains and in silage, but its main risk for human health is due to its presence in apples and apple juice[20] (Fig 3). Regulatory guidelines have therefore been set for its maximum concentration in apple juice at 50 μg l^{-1}.[21, 22] The sampling methods to be employed for foodstuffs have also been specified for the EU[23]. This specifies the taking of a composite sample with a mass of at least 1kg, aggregated from at least 10 increments for batches containing over 500kg.

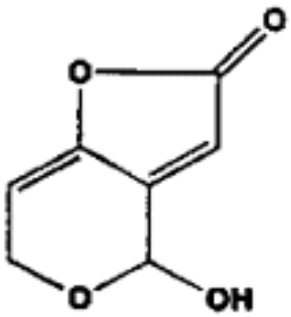

Fig 2. The chemical structure of patulin

Fig 3. Example of a glass of cloudy apple juice

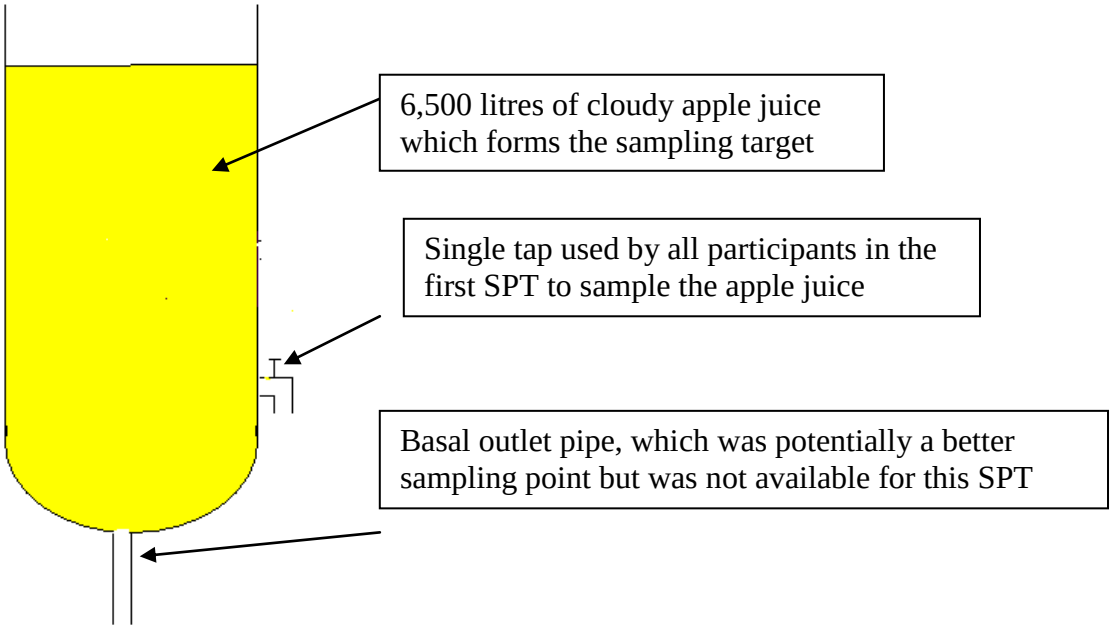

Fig 4. Schematic diagram of the sampling of cloudy apple juice from a process container containing 6500 litres, used in the first SPT

Cloudy apple juice is regularly monitored for patulin concentration during its production, and a large volume of freshly prepared juice was therefore considered as a potentially suitable sampling target for evaluation of SPTs. It was assessed in accordance with the 12 SPT Assessment Criteria[1] (and Table 1), for the first SPT of this research project, discussed in Section 4.4.

4.2 Experimental design for first SPT on food

The overall experimental design applied was the one generally used for SPTs, discussed above (Fig 1) and in the paper[1] (in Appendix G) but with a slight modification to its implementation.

The duplication of both the sampling and of the chemical analysis, for each of the m participants (where $m \geq 8$), allows the total variance from sampling to be separated into the three components between-sampler, within-sampler, and analytical. It also enables the performance scoring to allow for the heterogeneity of the target, and also to offset the effect of the analytical variance on each participant's score.

4.3 Methodology for first STP on food

1. Each of the nine samplers was given a sequential number to be used as an identifier.

2. The sampling target was a batch of 6500 l of unstirred cloudy apple juice contained in a metal process container (Fig 4). All 9 participants sampled from this one sampling target. Cloudy, rather than clear, apple juice was chosen for this experiment because it was expected to contain higher levels of patulin, and its cloudiness and unstirred condition were expected to be associated with higher levels of heterogeneity.

3. The SPT was designed and implemented in a way that minimised the influence that each sampler had on each other's interpretation of the sampling protocol. In particular the participants sampled at non-overlapping times (see Appendix B, Table B1 for timetable) and therefore did not witness each other's sampling technique.

4. On the day of the SPT (12/04/06) the samplers received a verbal briefing and written instructions before they started sampling (see Appendix C).

5. The samplers were instructed to try to sample in the same way that they usually did routinely, which could have been different for each sampler. The samples were all taken from the one tap connected to a hole in the tank, with a diameter of approximately 10 mm, situated approximately 1 m from the base of the tank (Fig 4).

6. Each sampler was instructed to draw two samples of the normal volume (194-251ml, Appendix F) directly into the usual sampling bottle (i.e. plastic 250 ml). It was requested that the second sample be a "new" sample, independent of the first sample. The samplers were instructed to wait approximately one or two minutes after completing the drawing of the first sample, before starting to draw the second sample, the objective being to make the two sampling events more independent within the sampling period.

7. The time schedule (Appendix B, Table B2) was used to record the time at which each sampler took their samples. The samplers were instructed to operate in numerical order with the minimum period between each sampler that would maintain their independence. Particular times for sampling were not specified in order to accommodate variations between the speeds at which the different sampler operated.

8. Immediately after a sampler had drawn a sample, it was given to the organizers who then labelled it (replacing existing labels if any) and put it into the laboratory fridge. On the label of each bottle the following information was recorded: the date, time and a sample code in the form SPT – N, where N represented a number that varied between one and the total number of samples drawn during the SPT (see Appendix D). Each sample was assigned a value of N, chosen (without replacement) from a

previously prepared list of random numbers (containing all possible values for N). Appendix D contains a list giving the correspondence between N and the number of the sampler, and whether it was the first or second sample drawn by that sampler. This design was used so that the lab analysis was not undertaken in the same order as the sampling, so as to minimize the effect of any drift in the analytical method.

9. After all the samplers had drawn their two samples each, the vessel containing the sampling target was emptied and sampled again periodically during that process for comparison against the SPT results to look for sampling bias. During this time, a total of five samples (of minimum 250 ml) were drawn at regular intervals, equally distributed over the entire time. The volume of juice remaining in the process container (from a gauge on side of the tank), and the time, was estimated at the moment each sample was taken. These samples were also assigned a random sample code and put into the fridge, together with the other routinely acquired samples (Appendix E).

10. Each sample was analyzed twice by the laboratory using a method conforming to the EU requirements for the determination of patulin concentration[23]. The lab was instructed that the sample must be shaken before an aliquot was taken for analysis. The same person analysed all of the samples over a period of 8 Days (Table 2). All samples were analysed in numerical order (N=1, 2, 3, etc., i.e. not in the order in which the samples were taken), with normal analytical quality control. The samples were analysed in duplicate, according to the experiment design, but not using exactly the same procedure. The two halves of the duplicate were analysed on different days, with the second half of the analytical duplicate being frozen to preserve it. This procedure was necessary due to lack of capacity in the laboratory to analyse such an usually large batch of samples.

11. The patulin concentration and the total sample volume in the bottle were determined for each sample. Other measurements were also made for: malic acid concentration, ascorbic acid concentration, sugar content, and Brix-Acid Ratio, but were they were not suitable for the SPT due to over-rounding (Appendix F).

12. The measurement results for patulin (e.g. Tables 2 & 3) were reported to the organizers in unrounded/untruncated format (i.e. raw values not edited for detection limits), as this is required for the accurate use of the statistical analysis.

13. Based on the analytical results, performance scores were calculated by the organizers for each sampler using the methods described in Section 4.6, and these were reported back to the samplers. The measurement uncertainty was estimated using the methods described in Section 4.5.

4.4 Assessment of the first SPT on patulin in apple juice according to the Criteria

The 12 general criteria derived in this project for the assessment of the effectiveness of an SPT (Table 1) have been discussed above and in the published paper[1]. The extent to which they have been met will be discussed here, to judge the effectiveness of this first SPT in the food sector.

1. **The size of the target must be large enough to allow sampling by at least 8 organisations, without significantly affecting the target.**

This criterion was met as the extraction of the all of the increments (plus spilled apple juice) had a cumulative volume of 5.043 l, which did not significantly influence the average composition of the target (volume 6500 l). The sampling ratio was therefore below the provisional value of 0.1%.

2. **Presence of an analyte that can determined with low enough analytical uncertainty to quantify sampling uncertainty.**

Probably met. Patulin was identified as probably being a suitable analyte. The reported measurement results range from 35.6 μg l^{-1} to 61.8 μg l^{-1} for the analysis of the increments extracted for the SPT by the samplers. All of the measurements were above the detection limit (approximately 10-15 μg l^{-1}). The proximity of the measurements to the detection limit probably explains why the analytical precision was relatively poor (~20%, at 95% confidence, see Section 4.5).

3. **Sufficient temporal stability of the sampling target (or else special procedures adopted to compensate for this)**

Probably met, because sampling was done quickly (the whole sampling experiment lasted no longer than 15 minutes) and the results do not show a significant trend with time.

4. **A level of heterogeneity of the analyte in the sampling target that is detectable with the proposed method of chemical analysis**

Not met. No significant heterogeneity was detected in the SPT results (see Section 4.5)

5. **The presence of a recommended sampling protocol applicable to the selected analyte**

Not met. We found no evidence of the existence of a written form of a sampling protocol, but only a verbal form based on communication between the staff. An EU agreed protocol for this purpose does exist[23], but it was not applied by this producer.

6. **Availability of appropriate samplers with the required technical skills**

Met. All samplers were readily available and fully trained in this task, but they probably lacked independence (see Criterion #10).

7. **Potential usefulness for improving performance of the samplers, and accessibility of the sampling target**

Not met. The sample drawing did not require much skill to implement the sampling procedure, and there was probably little or no room for improvement. This is partly caused by the homogeneity of the material, but also by the limited possible access to the target (only via a tap). However, there could possibly be improvements in their performance if the tap were relocated to be on the exit pipe at the bottom of the tank (Fig 4).

8. Required time to perform the sampling protocol and duration of the SPT

Met. The time to perform the sampling protocol (<3 min) and the duration of the SPT (<15 min) were very short compared to the rate of potential change in the composition of the sampling target, and also kept the cost of implementing the SPT to a minimum.

9. Acceptable level of cost due to damage to the target material and its packaging, during the SPT

Met. There was no damage to the batch as a whole, only a small loss of material (5 l), and only minor disruption to the production process, the cost of which was covered by the producer.

10. Independence of the samplers participating in the SPT

Probably not met, because all samplers were employed by the same company (i.e. the manufacturer of the apple juice used as the sampling target) and therefore they probably trained together to some degree.

11. Agreement of the owner of the sampling target and any sponsor of the SPT

Met. This agreement was obtained prior to the experiment.

12. Independence of the different contributors to the SPT

Not met. There was independence between the organisers of the SPT and the owner of the target (i.e. the manufacturer), but not between the owner of the target and the participant samplers (employees of the manufacturer), and the analytical laboratory (owned and on the same site as the manufacturer). There were therefore several potential conflicts of interest.

The broader interpretation of these findings will be discussed in Section 5.1.

4.5 Uncertainty estimation from SPT results

This discussion focuses on the estimation of uncertainty components from the results from the first SPT (for patulin in apple juice), based upon previous developed methodologies. The very similar method used for the second SPT (for moisture in butter) is given in the published paper[1](Appendix G).

As described in this paper[1], the evaluation of the measurement uncertainty using the data from this SPT in general is based on the output from the robust analysis of variance (RANOVA). Robust estimators are used to accommodate the effects of up to 10% of outlying values that do not conform to a Gaussian distribution[24]. The total variance is resolved into 3 components; analytical, within-sampler and between-sampler,

$$\sigma_{total}^2 = \sigma_{between\text{-}sampler}^2 + \sigma_{within\text{-}sampler}^2 + \sigma_{analytical}^2 \quad \ldots \text{Equation (1)}$$

Estimates of the variance of the population (σ) made experimentally (s) give the working relationship:

$$s_{total}^2 = s_{between\text{-}sampler}^2 + s_{within\text{-}sampler}^2 + s_{analytical}^2 \quad \text{…. Equation (2)}$$

The analytical variance is based upon the difference between the analytical duplicates (i.e. *m*.1.1 & *m*.1.2, and also *m*.2.1 & m.2.2, for sampler *m*, in Fig 1). It gives an estimate of the portion of the measurement uncertainty arising from the repeatability of the analytical method, but ignores the contribution from any analytical bias. The within-sampler variance comes from the duplicated samples taken by each sampler (i.e. m.1 and m.2 in Fig 1), after subtraction of the analytical variance. This is equivalent to the uncertainty estimate that would have been expected if the 'duplicate method' was applied to this one target using a single sampler, rather than to the minimum of 8 targets usually recommended[14]. The ANOVA separates the between-sampler variance as that which remains when the within-sampler and analytical variances are subtracted from the total variance. This variance quantifies the contribution to the uncertainty caused by factors such as the sampling bias between the participants. The experimental design in this case removes any contribution from analytical bias between the participants. This was achieved by using a randomized batch, and by centralizing the analyses in one laboratory in a randomised batch run under repeatable conditions.

The variance components from Equation 2 are quantified with RANOVA, and converted into the estimates of the corresponding measurement uncertainties. For this purpose the estimate of analytical repeatability, estimated across all of the participants' test materials, is included as part of the 'within-sampler' uncertainty.

$$U_{within\text{-}sampler} = 2 \times \sqrt{(s_{within\text{-}sampler}^2 + s_{analytical}^2)} \quad \text{….. Equation (3)}$$

To estimate the multi-sampler measurement uncertainty ($U_{multi\text{-}sampler}$) and standard deviation ($s_{multi\text{-}sampler}$), the between-sampler estimate ($s_{between\text{-}sampler}$) is added:

$$U_{multi\text{-}sampler} = 2 \times \sqrt{(s_{between\text{-}sampler}^2 + s_{within\text{-}sampler}^2 + s_{analytical}^2)} \quad \text{…. Equation (4)}$$

So in effect, using Equation 2

$$U_{multi\text{-}sampler} = 2 \times s_{total} \quad \text{…. Equation (5)}$$

and

$$U_{between\text{-}sampler} = 2 \times s_{between\text{-}sampler} \quad \text{…. Equation (6)}$$

4.5.1 Uncertainty estimates from the first SPT on patulin in apple juice

The method of evaluation used is basically the same as the general one described above and used for the second SPT[1]. A slight modification was required because the experimental design (Fig 1) was slightly changed in practice. The sample preparation for the two halves of the analytical duplicate for the determination of patulin were conducted differently, with the second half of the duplicate being frozen before analysis, for logistical reasons with the laboratory. The measurement results for the first duplicate analyses are given in the column with header "Dup 1" (Appendix F) and the results for the second duplicate analysis under the header "Dup 2". A statistically significant bias (of 22%) was detected between these duplicate values, with lower values being reported for the second duplicate (that had been frozen), as might be expected. Only the first duplicate measurement (Table 2) was therefore

used for statistical analysis, in what was effectively a simplified experimental design[15, page 98].

Table 2. Concentrations of patulin (µg l^{-1}) measured once (on the first analysis, A1) on both of the duplicates samples (S1 & S2) taken by the 9 participants in the Apple Juice SPT. Visual inspections suggests that there is as much variability within–sampler as there is between-sampler, a conclusion that is substantiated by the ANOVA results.

Sampler	S1A1	S2A1
1	54.7	48.7
2	49	48.1
3	46.1	50.5
4	52.1	56.4
5	52.2	61.8
6	51.6	56.2
7	46.3	56.6
8	52	58.4
9	50.2	55.6

The multi-sampler expanded measurement uncertainty (i.e. for sampling and analysis combined), estimated with robust ANOVA [25], is 19.4%, (i.e. total relative standard deviation is 9.7%). No significant between-sampler variance was detected, so in this case the SPT approach does not produce a different estimate of uncertainty to that produced by the 'duplicate method' using the within-sampler variance (addressing Objective 4). The lack of effective analytical duplicates means that measurement standard deviation (and hence uncertainty) cannot be separated into the contributions from the sampling and the chemical analysis by this means. However, the laboratory's Internal Quality Control Scheme did analyse a daily check sample, subject to a uniform preparation between batches. These results are given in Table 3. The relative standard deviation of the check samples over both runs is 9.8%. This can be used as a very crude external estimate for the analytical repeatability precision, which can also be used as a very approximate estimate of the analytical uncertainty.

Table 3. Internal Quality Control Results for the patulin in a 'check sample' of apple juice analysed in the batches with the SPT samples

	patulin in check sample (µg l^{-1})
Check sample results for batches with **1st** half of sample duplicates (run fresh without freezing)	
Day 1 = spt 1 – 7	26.8
Day 2 = spt 2 – 17	28
Day 3 = spt 18 – 23	24.4
Check sample results for batches with **2nd** half of sample duplicates (run after freezing)	
Day 4 = spt 1	24.4
Day 5 = spt 2 – 5	27.1
Day 6 = spt 6 – 9	24.5
Day 7 = spt 10 – 14	26.2
Day 8 = spt 15 – 16	32.2

Taking into account the low numbers of measurements, there is no significant difference between the estimates for analytical precision from the check sample (9.8% as RSD, or 20% at 95% confidence) and the estimate of the measurement standard deviation (9.7%) from the SPT materials. This similarity would suggest that the contribution to the uncertainty from the sampling is very low. This indicates that the heterogeneity of the sampling target is also very low, and probably insufficient for undertaking an effective SPT (Criterion # 4, Table 1). The probable dominance of the analytical uncertainty (U_{anal} = 20%) suggests that the analytical method used was not well suited for conducting this SPT, probably due to the proximity of the typical analyte concentration (46-62 μg l^{-1}) to the detection limit of the method (approximately 10-15 μg l^{-1}, see Criterion #2).

Possiblility of sampling bias in the SPT.

The possibility of overall sampling bias was investigated by taking 5 apple juice samples from the same side tap on the tank while it emptied after the completion of the SPT. The mean patulin concentration of these samples (mean 51.9, SD 4.96 μg l^{-1}, Appendix F) does not show any significant differences from the mean value from the SPT (classical mean 52.6, SD 4.50 μg l^{-1}). This finding does not, however, preclude the possibility that all of the samples taken from the one tap on the side of the tank might still be biased compared to the true value. If the particulate matter in the unstirred tank tended to fall to the base of the tank, and the patulin preferentially adheres to the particulate matter (which has been reported[26]), then it is probable that the higher concentrations of patulin at the base of the tank was never sampled. A more representative sample could therefore probably be obtained by sampling the outflow pipe at the base of the tank (Fig 4). In a future study, samples taken at that point could be used to check for an overall sampling bias that would affect all of the samples in an SPT equally, and would therefore not be detectable by the current experimental design.

4.6 Calculation of performance scores in SPTs

There are many possible ways of calculating a performance score for PTs, and the differences in requirement for scoring APTs and SPTs have been considered[1]. The general z-score for analytical proficiency testing is defined by the International Harmonized Protocol[27,2].

$$z = (x-X)/\sigma_{target} \quad \text{....Equation (7)}$$

where z represents the reported z-score, x is the participant's result, X is the assigned value and σ_{target} is the target value for the standard deviation. It is noted that the International Harmonized Protocol is intended to be applicable for the proficiency testing of analytical chemistry laboratories (APTs). The method of scoring needs to be somewhat different for a Sampling Proficiency Test (SPT), but the general concept of a z-score has been shown to be applicable to SPTs as well, and has been described for environmental SPTs[7,8,9,11].

There are several options possible for representing the variables x, X, and σ_{target} in Equation 7 for both APT's and SPT's. Two simple options for the choice of X are the classical mean and the robust mean of all results. For x and σ_{target} a range of options have been investigated (Table 4). The options for x are denoted by A & B and those for σ_{target} by I &II.

The first possible option that is discussed is denoted as Option AI. For the target standard deviation it is based upon an externally set "fitness-for-purpose" criterion (Option I). For

the participant's measurement of concentration (x), the average of both sample results was used (Option A). The scores calculated for the prototype SPTs using Option AI can be found below (Table 5).

The second option (B) for selecting a value for x is to use the individual analytical measurements on both duplicated samples, instead of the average (Option A). The Option BI, with the robust mean for X, was applied to the second SPT which measured the moisture content of butter[1].

Table 4. Possible scoring options for SPT's explained in terms of codes for each option for both the participant's result (x) and the target standard deviation (σ_{target})

Option Code for x	**Option for x**		**Option Code for σ_{target}**	**Option for σ_{target}**
A	The average analytical result on both samples		I	An externally set fitness for purpose standard deviation
B	Individual analytical results on both duplicate samples		II	The total classical standard deviation of measurement

4.6.1 Calculation of performance scores for first SPT on patulin in apple juice

Only one patulin measurements (Dup 1 in Appendix F) was used for the scoring of the participants, for the reasons explained above. The patulin concentrations were the only ones used for scoring, as they are the most important analyte to the manufacturer, and they also showed much more variation than those for Brix, Malic Acid and the Brix/Acid ratio, which were all over-rounded.

In order to be able to use one scoring Option AI, a value for the fitness-for-purpose standard deviation was required. Unlike the case for an APT, the value σ_{target} must allow for the heterogeneity of the sampling target, and the uncertainty from the chemical analysis, so that samplers are not penalized for these factors which are beyond their control. The fitness-for-purpose value for this standard deviation was unknown to the producer of the juice, and therefore to the SPT organisers. Various possible values for this parameter were studied, therefore, to gauge their effect on the resultant z-scores, for all nine samplers (Table 5).

Table 5. Participant's z-scores for the samplers in the Apple Juice SPT as a function of the selected fitness-for-purpose standard deviation (FFP-SD), using Option A1 (see Table 4). Scores with absolute values larger than +2 are highlighted in bold. All scores are proficient at FFP-SD levels at and above 10%, indicating that no non-proficient samplers were detected (see text).

FFP-SD[(i)] (σ_{target}) → Sampler ↓	5%	10%	15%	20%
1	–0.47	–0.23	–0.16	–0.12
2	**–2.16**	–1.08	–0.72	–0.54
3	**–2.31**	–1.15	–0.77	–0.58
4	0.90	0.45	0.30	0.22
5	**2.36**	1.18	0.79	0.59
6	0.71	0.35	0.24	0.18
7	–0.59	–0.30	–0.20	–0.15
8	1.39	0.69	0.46	0.35
9	0.17	0.09	0.06	0.04

(i) Fitness-for-purpose standard deviation is expressed as a percentage of the (classical) mean of the data.

If it is assumed that a result is not fit-for-purpose when the absolute value of the performance score is larger than 2, then non-proficient samplers arise only when the FFP-SD (σ_{target}) is set at 5%. As the value of the FFP-SD increases the absolute value of the z-score decreases, as would be expected from Equation 7. (i.e. all samplers are classified as proficient).

The within-sampler uncertainty (9.7%), which includes any contribution from the analytical precision (Equation 3), can be used as an estimate of the heterogeneity of the sampling target. This would suggest that σ_{target} must not be less than 9.7%, and potentially at least $1.5 \times 9.7\% \approx 15\%$, to show that the performance of a sampler is not fit-for-purpose. From Table 5, it can be seen that all of the samplers gave results that are fit-for-purpose for $\sigma_{target} = 15\%$.

The limitations of this first SPT, that have already been noted as insufficient heterogeneity and relatively poor analytical precision, also have the effect of not enabling effective scores of the participants to be calculated. These limitations were addressed in the second SPT, which then did enable effective scores to be calculated and communicated to the participants.

5. Discussion

Issues identified in the first SPT helped to improve the design of the second SPT, and the results from both will help in the overall assessment of SPTs in the food sector in general.

5.1 Discussion of findings from the first SPT

This very first SPT in the food sector proved the feasibility of the approach, but highlighted several issues that need careful attention in the design and interpretation of later SPTs.

Overall the first SPT met, or probably met, most of the assessment criteria, but five main limitations were identified.

5.1.1 There was insufficient heterogeneity in the sampling target (Criteria 4, Table 1). This was evident because the overall uncertainty was dominated by uncertainty arising from the analytical process. The situation may have been exacerbated by the relatively high analytical precision (20% at 95% confidence), due to the analytical detection limit being too close to the typical concentration values in the samples (3 to 6 times, instead of the ideal 10 times). However the main issue is that unless a sampling target is sufficiently heterogeneous, then differences in proficiency between samplers won't be detectable, and no effective scores can be calculated.

5.1.2 There was a lack of independence, both between the samplers (Criteria 10), and between the owner of the sampling target and the analytical laboratory (Criteria 12). The degree of independence of samplers would probably be improved by getting samplers from different organizations, but there may then be issues of commercial confidentiality for sampling targets such as the these stored within the premises of one company. The lack of independence between the owner of the sampling target and the analytical laboratory could have affected the performance of the in-house laboratory. It could be easily overcome, with foresight and sufficient resources, by sending all of the SPT samples to an independent lab that has expertise in this suite of chemical analyses. An alternative, which has been tried in the environmental sector, is for each participant to undertake their own chemical analysis, with a matched reference material in common to monitor between lab-bias[7].

5.1.3 There was poor physical access to the material in the target (Criteria 7), which was only available via one poorly located tap. It would have been preferable, therefore, to get a more representative cross-section of the whole target over all depths within the tank by using a tap on the outflow tube at the base of the vessel, if this had existed (Fig 4). The more general point is that access issues can limit the performance of the samplers. The study demonstrates that choosing a sampling target with flexibility in samplers' access is crucial for an effective SPT.

5.1.4 There was a lack of effective analytical duplicates in the laboratory, the use of which is specified in the general experimental design for an SPT (Fig 1). In this case, this was due to by unforeseen logistical issues (lack of space for all of the SPT samples in the lab fridge). It was not possible, therefore, to separate the analytical component of the uncertainty based upon this design, but it was possible, if not very accurate, to use an external estimate of this statistic. The general issue is to ensure in advance that there is sufficient capability and capacity in the chemical laboratory used in the SPT. This would also extend to making sure that the lab randomized the samples within the analytical batch, and reported the measurements in an unrounded and un-truncated format.

5.1.5 The sampling protocol applied was apparently only based upon verbal communication, with no written version being available (as required by Criterion #5), despite the existence of an agreed EU protocol[23].

Overcoming these limitation for future SPTs, and ensuring they conform to the rest of the 12 criteria (Table 1), should ensure that the calculation of the participants' scores, and the estimates of uncertainty, will not be adversely effected.

5.2 Lessons learnt from the first SPT, that were applied to the second SPT on moisture in butter

The findings of the second SPT, on moisture in butter, are discussed in detail in the published paper[1] (Appendix G). The relationship to the findings of the first SPT, and the way these improved the experimental design of the second SPT, will be discussed here.

The second SPT aimed to overcome as many of the five limitations identified in the first SPT as possible, within the logistical constraints.

1. The distribution of the measurand in the sampling target (moisture in butter), was expected not to be very heterogeneous from a previous study on frozen butter which gave a within-sampler value of 0.19% m/m RSD[28]. This proved to be the case, with an experimental value in this study of 0.08% m/m RSD. However, the precision of the analytical method for the gravimetric determination of water was known to be very good (later estimated to be 0.17% RSD). This meant that although the analytical precision was not ideal for accurately characterizing this heterogeneity (i.e. a tenth of the heterogeneity = 0.008% RSD) there was considered to be a reasonable chance of meeting Criterion #4. In the event there proved to be sufficient heterogeneity to quantify the between-sampler effect (0.78% relative) and to detect two non-proficient samplers[1].
2. To address the second limitation, the samplers were recruited mainly from different local authorities. There was, however, still some potential lack of independence caused by some levels of previous joint training. The importance of the 'availability of appropriate samplers with the required technical skill' (Criteria #6, Table 1) was tested by including one untrained sampler. The independence of the lab was ensured by using an external, independent laboratory, which analyzed the samples under randomized repeatability conditions.
3. The third limitation of needing flexible access to the sampling target was much easier to achieve for the 20.1 tons of butter than it was for the apple juice. The fresh butter was stored in 804 separate boxes, each containing a 25kg block, and located on numerous pallets on the floor of a cold room, with access on four sides to meet Criterion #7 (Fig 5).

Fig 5. A sampler gaining free access to any of the 804 boxes that contained the sampling target of 20.1 tons of fresh butter

4. The requirement to get effective analytical duplicated measurements was achieved by the selection of a suitable laboratory, and ensuring that this was provided in advance with a full briefing of the analytical requirements and the batch size.
5. To meet Criterion #5, it was established that a written sampling protocol[29] was available to be applied in the second SPT.

The second SPT did detect between-sampler bias, made improved estimates of the measurement uncertainty to include this, and did identify non-proficient samplers[1], whereas the first SPT did not. This was probably due to overcoming the lack of heterogeneity in the sampling target, and the relatively poor analytical precision in the first SPT.

5.3 General issues of feasibility, practicality and usefulness

The overall aim of this research was to assess the feasibility, practicality and usefulness of SPTs in the food sector. The approach that was taken to address this aim was mainly the implementation of the one initial or prototype SPT (patulin in apple juice) followed by second SPT (moisture in butter) with an improved design.

The *feasibility* of designing and implementing SPTs has been proven. Both of the SPTs undertaken in this research project showed how all of the necessary components can be brought together effectively. The first SPT was a necessary part of a learning process, in which the selection criteria were identified, based largely upon the experience gained during the test. These 12 criteria encapsulate the key aspects that need to be considered in advance, in order to make an SPT both *practicable* and effective. It did prove possible to find owners of potential sampling targets that would be willing to make available the material and buildings required to hold the two SPTs. However, it was noted that there was a significant level of reluctance in many organisations in the food industry to host or even take part in SPTs. Finding ways to improve the motivation of potential SPT participants is therefore an important future task.

These criteria also relate to the assessing the potential *usefulness* of SPTs in the food sector. The purpose of the first SPT, as with all prototypes, was not to be perfectly effective initially, but to help to identify improvements that need to be made so that greater effectiveness can be achieved in the longer term. For example, the very first prototype environmental SPT[7] proved the logistical feasibility of SPTs generally in that sector, but was not perfectly effective in several respects. It raised questions, such as how the 'target value of concentration (X)' should be set, and whether the spatial location of the contamination should also be reported, rather than just the mean value. This was partially resolved by a later SPT, in which the spatial information and target value were created synthetically with known values, to enable more robust scoring of the SPT[11]. The practical usefulness was therefore limited in the first prototype, but substantially improved the effectiveness of later SPTs as a result of subsequent improvements.

By analogy, these first 'prototype' SPTs in the food sector illustrated the particular requirements and limitations of applying these ideas to this new sector. The usefulness of the results from both the first, and to a lesser extent the second, SPT can be summarised as:-

1. It has been shown that SPTs can be conducted in the food sector, on both liquid and solid materials.
2. A series of 12 assessment criteria can usefully be applied, both before and after an SPT, to decide whether the SPT is likely to be effective.
3. Sufficient heterogeneity (Criterion #4) is a more important requirement of the sampling target to meet in the food sector, than in the environmental sector where it is more often present. The first SPT lacked sufficient heterogeneity of the key analyte in the target, meaning that it was not possible to detect differences between the proficiency of samplers using typical sampling protocols. This could also be

viewed as a positive attribute for the food industry, because it shows that the quality of the sampling using those protocols is not particularly sensitive to the particular skills of the samplers.

4. Where the precision of the analytical method is poor (e.g. 20% for patulin in apple juice in the first SPT) the power of the SPT to detect differences between the samplers can be lost. This limitation was overcome in the second SPT (analytical precision 0.4% for moisture in butter), but it also depends on how the precision compares against the heterogeneity of the target.
5. An SPT in the food sector has been shown to be effective in detecting non-proficient samplers. It showed high z-scores (> +/- 2), and rescaled sum of z-scores (> +/- 3), for two samplers participating in the Butter-SPT (one untrained, but another a trained professional)
6. In the process of organising SPTs in the food sector, it became evident that there was some un-willingness of organisations to participate, both as a participant, and as the host of an SPT. The issue of improving this motivation, by explaining the advantages of SPTs, requires future consideration.
7. Nearly all of the professional samplers in both of the prototype SPTs performed acceptably, according to their z-scores. This could be seen as an ideal outcome for a routine SPT, as it shows that very few of these professional samplers (i.e. one) were non-proficient. It was not, however, ideal for a prototype SPT where a demonstration of the ability to detect more non-proficient samplers would have been informative.
8. Another cause for concern is that all samplers in both SPTs chose the same sampling protocol, making these IOSTs more like collaborative trials in sampling (CTS). In previous environmental SPTs, the freedom of samplers to choose their own sampling protocols, based on their professional judgement, was part of the test of their proficiency. The variation in the protocols chosen, and their differing interpretations even within one protocol, may well have contributed to a much wider range of z-scores[11]. An alternative interpretation is that in the food sector the samplers are better trained, and always do choose the correct protocol for a given circumstance, but further food-based SPTs would be needed to verify this.
9. The poor definition of a sampling target[1] was shown to cause problems in the estimation of measurement uncertainty and in the scoring of participants.
10. SPT results can be used to estimate total measurement uncertainty, including not only sampling precision, but also the potential bias between different samplers. For example, in the Butter SPT the uncertainty estimated using the SPT results (0.87%, relative, for 95% confidence) was twice the value of that estimated using single sampler methodology (0.39%)[1]. The latter method is equivalent to the 'duplicate method' that is recommended[15] and often used for this purpose, so the SPT value may well be providing a more reliable estimate of the uncertainty.
11. The best experimental design for an SPT in the food sector is probably that used here (Fig 1). Although it costs more to implement both duplicate analyses and duplicate samples for each participant, it has the great advantage of enabling the effect of sample heterogeneity to be quantified (in the within-sampler variance). It also allows the effect of the analytical precision on the outcomes (both z-score and uncertainty) to be quantified (as the analytical variance). This conclusion is certainly valid for prototype SPTs, such as those described here, but it may also be that in a routine SPT on a well characterised sampling target, that simplified designs might be evaluated in future (e.g. an unbalanced design[15, Page 98]).
12. The best method for scoring an SPT is still not certain, but one method is a possible candidate. In the best experimental design (Fig 1) each participant produces four results, so the options of scoring by using just one of these values (Option A1)

seems not to make best use of all of the information. The use of averages of some, or all, of these four numbers, will tend to smooth out the effects of individual measurement results. One promising option is to calculate individual z-scores for each result, but then use some sort of aggregate score based on all four numbers. One approach to this would be to use the rescaled sum of the z-scores[30] (RSZ = $\sum z/\sqrt{m}$), but another option, or a different threshold, may be required to overcome the effect of the lack of independence between the four z-scores for each participant.

5.4 Suggestions to improve the usefulness of future SPTs in the food sector

1. ***Perform a further SPT*** on a sampling target/ analyte combination that more closely fulfils the 12 assessment criteria for an effective SPT (Table 1). One example could be the sampling of a large batch of groundnuts for aflatoxins. Initial investigations during this project suggested that such a typical 20 ton batch at a sea port would be expected to have sufficient heterogeneity to provide a sensitive sampling target for detecting differences between the proficiency of different samplers. Recruitment of participants should aim to attract samplers from different ports that undertake this sampling, but which are geographically well separated, so as to get more independence between the samplers. They should also be encouraged to choose whatever sampling protocol they consider appropriate according to their professional opinion, rather than simply being steered into applying the same protocol (as happened in both of the SPTs in the current study).

2. ***The definition of the sampling target should be unambiguous***, and correspond exactly with that used by the vendor of the material for trading purposes. This should ensure that the target standard deviation of scoring the SPT can be based upon the fitness-for-purpose specified by the vendor (if it is available).

3. ***Separate the roles of ownership of the sampling target, SPT participant, and analytical laboratory.*** None of the participants should own the sampling target or the analytical laboratory, and therefore should have no potential conflicts of interest concerning the implications of the findings to any management decisions that are made on the quality and fate of the target material. The analytical lab, in an SPT of this design with centralised chemical analysis, should not be one of the SPT participants or the owner of the sampling target (unless that is the organiser of the SPT, which is perhaps the ideal situation).

4. ***Location the sampling target.*** It would be preferable to locate the sampling target for the SPT in a location that would enable samplers from different organisations to participate without any concern about the commercial confidentiality of a production process. From this perspective a neutral location would be preferable to one owned by one of the participants.

5. ***Improve the motivation of organisations and samplers to participate in SPTs.*** This might be achieved by requiring participation in SPTs for the accreditation of sampling (or the certification of samplers). This approach has clearly increased participation of laboratories in APTs, which are now well established in the food, and many other sectors. In the early development of APTs only the lab, rather than the whole organisation, was involved in what were then called ‘round robins’. As a result, informal participation in PTs was not often resisted by company management. By contrast, SPTs require the cooperation of the senior management of a company right from the start. Moreover, the sampling personnel are often

managed from outside the established measurement quality systems of food production organisations, and their training and supervision often seem to be less rigorous than they are for laboratory staff. It is perhaps not surprising that although measurement scientists in many organisations are often initially enthusiastic about taking part in an SPT, the senior management is usually less so. It is particularly important to find ways to motivate organisations who are sampling heterogeneous material using minimally trained samplers, and which can therefore potentially benefit most from participating in a series of SPTs.

6. ***There could be some advantages in holding a Measurement Proficiency Test (MPT) rather than an SPT.*** In an MPT each participant would not only take their own sample, but also conduct their own sample storage, physical preparation and chemical analysis. This would simplify the performance scoring, as no allowance would have to be made for separating the effects of sampling and analysis. It would not matter in this case whether the analytical method was a large source of uncertainty (Criterion #2 less important), as it would be the overall performance, rather than just that of the sampling, that would be assessed. It would be up to the participant to decide the cause of any non-proficient score. In one published example of an SPT that was very similar to an MPT[7], a matched reference material was also distributed to all participants, in order to enable some level of diagnosis for non-proficient scores.

7. ***Steps to reduce atypical sampling behaviour in an SPT.*** The proficient z-scores of nearly all of the professional samplers in both prototype SPTs suggest that they may have lacked independence, and/or that they may have been trained intensively before the SPT took place. This raises an important issue, which also arises in APTs, that the performance of participants in a formal PT may not well represent that which is typical of routine operations. It future it would be better to recruit the samplers that are more independent of the target owner, and only give them exact details of the sampling target close to the time of the SPT. Such a procedure was employed in an environmental SPT[11], and a greater proportion of non-proficient scores was recorded (33% $|z|>2$, 22% $|z| >4$). Perhaps these arrangements for a food SPT would result in scores that more accurately reflect the performance of the samplers under routine conditions.

8. ***Organiser of SPTs should be encouraged to try to quantify any overall sampling bias that may be affecting all of the participants.*** The first SPT on patulin in apple juice illustrated that although all of the participants agreed with each other on the patulin concentration, they may have all been equally biased. This was due, in this case, to the fixed location of the only sampling point on the side (rather than the base) of the tank used to hold the batch of apple juice. The SPT organisers attempted to estimate any possible bias, by sampling the batch as it flowed out of the tank, but this again had to be taken from the tap on the side, rather than one that could have been placed in the outflow pipe at the base of the tank. This example does, however, illustrate the general principle of SPT organizers aiming to take an independent and more representative sample of the batch, by a non-routine protocol, in order to quantify any sampling bias. This bias could then be included in the report to participants and included in estimates of the overall measurement uncertainty[15].

6. Conclusions and further work required

This research achieved its aim by proving the feasibility of Sampling Proficiency Tests (SPTs) in the food sector by selecting and undertaking two SPTs (patulin in apple juice and moisture in butter). The first SPT, as a prototype, was not intended to be perfect, but was used to help devise and refine 12 criteria (Table 1) that can be used to assess whether the results of such SPTs are likely to be effective in providing samplers with scores that truly reflect their proficiency. When these 12 criteria were applied to the first SPT several limitations were identified, including

- a lack of sufficient heterogeneity of the key analyte in both sample targets
- poor analytical precision that obscured differences between the samplers
- lack of independence between the samplers
- poor access to the material in the target

These limitations were largely overcome in the second SPT, but this did have a different limitation:

- poor definition of the sampling target (the batch of butter) which complicated the scoring of participants, and the estimation of uncertainty

Most of these deficiencies can be overcome in future SPTs by more careful selection of the sampling target, the key analyte, and the participant samplers. In terms of the more general practicality of setting up SPTs, a key issue is a need to improve the motivation of organisations in the food sector to participate. One approach to this may be to work toward the accreditation of food samplers by UKAS, and to include the participation in SPTs as a requirement of that accreditation. Accreditation of samplers is already underway in some other sectors, such as the monitoring of gaseous emissions from stacks, and is likely to be extended[31]. Participation in analytical proficiency tests (APTs) is already a requirement for laboratory accreditation under ISO 17025.

The ultimate uptake of SPTs in the food sector may well be driven by their perceived usefulness. These prototype SPTs have been very effective in identifying how SPT design needs to improve, and how their usefulness can be assessed.

It did prove possible to devise a scoring scheme for SPTs that took into account two differences from APTs. Allowance was made for the necessary heterogeneity of the sampling target, and also excessive analytical uncertainty that could potentially cause outlying scores unrelated to the proficiency of the sampler. The best experimental design (Fig 1), which was fully implemented in the second SPT, enabled the uncertainty caused by the heterogeneity and the analysis to be quantified, as within-sampler and analytical variance respectively. The between-sampler effect, which is essential to an effective SPT, was thereby separated and quantified. It may be that performance scores should not be reported if there is no detectable between-sampler variance, as they are unlikely to reflect the performance of samplers. For the apple juice SPT no between-sampler effects were detected, probably due to a combination of insufficient heterogeneity in the target and poor analytical precision. For the butter SPT, evidence for two non-proficient samplers was detected. One was a non-professional sampler purposefully included in the SPT to see if a difference in performance could be detected. The other borderline non-proficient sampler was a professional who had a single z-score above 2. The potentially best scoring scheme, based on the best experimental design, generates 4 z-scores per participant, but combines them in one value (with a technique such as a rescaled sum of z-scores ($RSZ = \sum z/\sqrt{m}$). More research is required to refine this scoring method, so that it is shown to fairly reflect the proficiency of samplers, over a wider range of sampling targets.

It also proved possible to use the SPT results to estimate the uncertainty of measurements of concentration, by a quite novel approach. European Guidance[15] suggested making such estimates of uncertainty from sampling (and analysis) using a single sampler approach, using what is called the 'duplicate method'. This research showed that using the multiple samplers in the butter SPT, the estimate uncertainty (0.87%) was substantially higher than that from the single sampler approach (0.39%). This is because of the contribution of the 'between-sampler' effects, which probably reflects bias between their sampling techniques. Such a difference was not detected for the apple juice SPT, probably due to the insufficient heterogeneity and/or the relatively poor analytical precision.

Further research is required on a wide range of topics within this new field, if the full potential of SPTs in the food sector is to be realised (discussed more fully in Section 4.4).

Proposals include:-

1. The holding of a further SPT on a food target, with sufficient heterogeneity and more independent samplers, that has the potential to show the different proficiencies of the participant samplers more clearer. It will require:-
   - an unambiguous definition of the sampling target.
   - separated roles for the owner of the sampling target, and the SPT participants.
   - an independent location for the sampling target, that encourages participation of participants from a wide range of organisations.
   - an independent analytical lab.
   - assessment of the SPT using the 12 criteria to be met for an effective SPT (Table 1).
   - steps to be taken to make the sampling behaviour in the SPT more typical of that used in routine sampling
   - mechanisms in place to quantify any overall sampling bias that may be affecting all of the participants (e.g. by independent sampling of the batch using another technique).
2. Aim to improve the motivation of organisations and samplers to participate in SPTs, perhaps by collaboration with UKAS on the accreditation of sampling.
3. Consider holding a Measurement Proficiency Test (MPT), rather than just an SPT, in which the participants are responsible for the whole measurement process (i.e. sample preparation and chemical analysis, as well as sampling). This will more accurately reflect the performance of whole organisations and the reliability, and total uncertainty, of the resultant measurements.

Appendices

Appendix A: Glossary of terms and abbreviations

This report contains technical terms and abbreviations, which are defined here so as to assist the reader.

ANOVA	Analysis of variance
APT	Analytical proficiency test
CTS	Collaborative trial in sampling
FFP	Fitness for purpose
IOST	Inter-organizational sampling trial
ISO	International organization for standards
PT	Proficiency test
RANOVA	Robust analysis of variance
SD	Standard deviation
SNF	Solids non fats
SPT	Sampling proficiency test

Sampling Process of drawing or constituting a sample.
ISO 11074-2 (1998)[32], ISO 3534-1 (1993)[33]

Sampling target Portion of material, at a particular time, that the sample is intended to represent.
AMC (2005)[34]

Uncertainty from sampling The part of the total measurement uncertainty attributable to sampling.
Note. Also called sampling uncertainty
IUPAC (2005)[35]

A fuller glossary of terms relating to sampling, and the measurement uncertainty it generates, can be found elsewhere[15].

Appendix B: Time schedule of sampling: first SPT on patulin in apple juice

Table B.1. Schematic sampling schedule, with the actual times of sampling noted

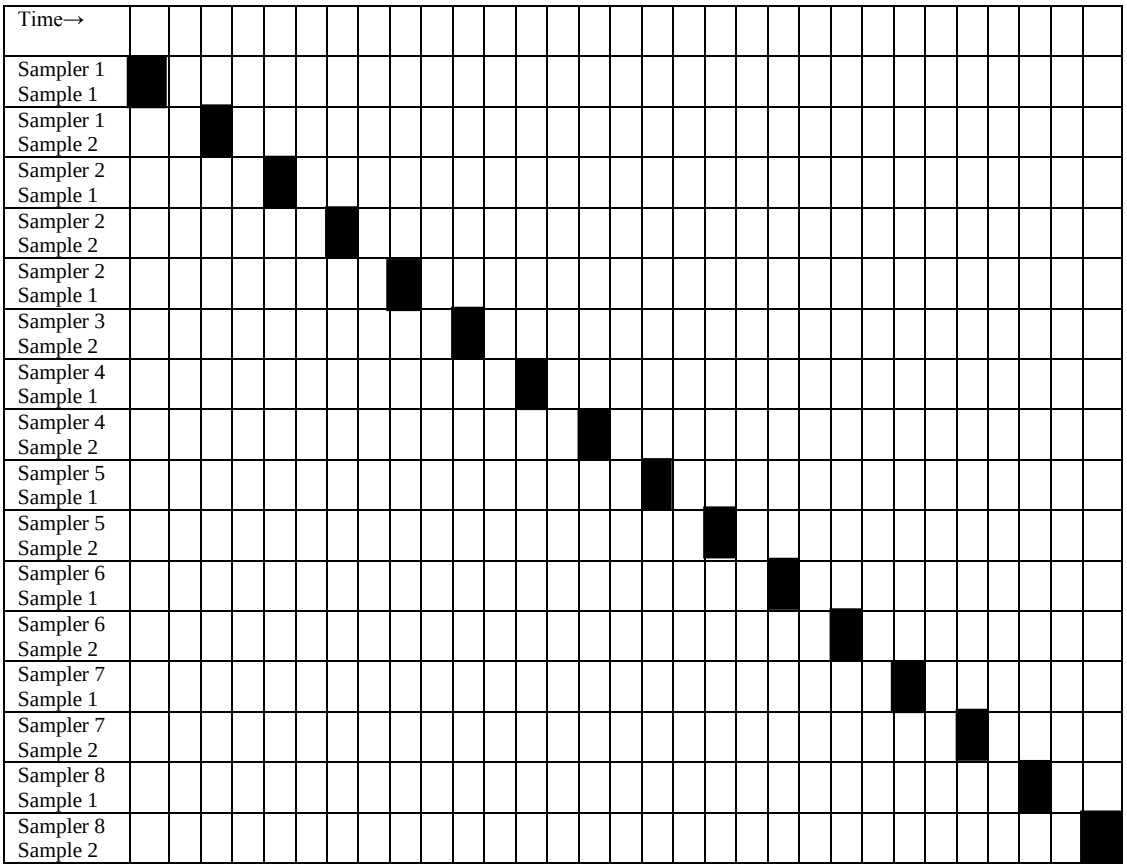

Table B.2. The actual times of sampling for first SPT on patulin in apple juice

Sampler	Time
1.	3:16-3:18
2.	3:20-3:21
3.	3:22-3:23
4.	3:23-3:24
5.	3:24-3:25
6.	3:25-3:26
7.	3:27-3:28
8.	3:29-3:30
9.	3:30-3:31

Appendix C: Instruction to samplers: first SPT on patulin in apple juice

Instruction for individual samplers

1. Two samples are to be drawn from the designated container of apple juice.

2. Both samples should be drawn in the same way as you would do routinely.

3. Between drawing the first and second sample, please wait approximately one or two minutes.

4. Clearly mark the samples with your number followed by “1” or “2” for the first and second sample respectively (e.g. 3/2 for the second sample taken by the third sampler).

5. Please give both samples to the organizers.

Appendix D: Random sample codes for first SPT on patulin in apple juice

Sample codes generated using random numbers.

Sample sequence number	Sampler	Duplicate sample	Sample code
1	1	1	SPT-6
2		2	SPT-3
3	2	1	SPT-1
4		2	SPT-16
5	3	1	SPT-14
6		2	SPT-15
7	4	1	SPT-10
8		2	SPT-5
9	5	1	SPT-22
10		2	SPT-4
11	6	1	SPT-8
12		2	SPT-21
13	7	1	SPT-17
14		2	SPT-12
15	8	1	SPT-23
16		2	SPT-19
17	9	1	SPT-18
18		2	SPT-2
19	Outflow 6500 l		SPT-7
20	Outflow 5000 l		SPT-20
21	Outflow 3500 l		SPT-13
22	Outflow 2000 l		SPT-9
23	Outflow 1200 l		SPT-11

Appendix E: Characteristics of discharged apple juice after the first SPT

Volume in tank (l, approx.)	**Sample code**	**Time**	**Temp (^{o}C)**	**Patulin conc.** ($\mu g\ l^{-1}$) (1^{st} Dup)
6500	SPT-7	16:40	8.7	57.0
5000	SPT-20	16:46	9.5	56.5
3500	SPT-13	16:47	9.0	45.1
2000	SPT-9	16:51	9.0	51.0
1200	SPT-11	16:54	9.5	49.9

Appendix F: Analytical results for the first SPT on patulin in apple juice

Sample Code	Patulin (ppb)[1],[3]		Brix (%)[1]		Malic Acid (%)[1]		Vitamin C (mg/l)[1]		Ratio[1],[2],[3]		Volume (ml)
	Dup 1	Dup 2	Dup 1	Dup 2	Dup 1	Dup 2	Dup 1	Dup 2	Dup 1	Dup 2	
SPT-1	49.0	44.5	12.8	12.8	0.42	0.42	200	200	30.5	30.5	204
SPT-2	55.6	46.6	12.8	12.8	0.43	0.42	200	200	29.8	30.5	224
SPT-3	48.7	40.3	12.8	12.8	0.43	0.43	200	200	29.8	29.8	243
SPT-4	61.8	47.3	12.8	12.8	0.42	0.42	200	200	30.5	30.5	196
SPT-5	56.4	38.9	12.8	12.8	0.43	0.43	200	200	29.8	29.8	210
SPT-6	54.7	39.6	12.8	12.8	0.43	0.42	200	200	29.8	30.5	243
SPT-7	57.0	40.0	12.8	12.8	0.43	0.43	200	200	29.8	29.8	228
SPT-8	51.6	42.9	12.8	12.8	0.43	0.43	200	200	29.8	29.8	206
SPT-9	51.0	41.5	12.8	12.8	0.43	0.43	200	200	29.8	29.8	235
SPT-10	52.1	42.8	12.8	12.8	0.44	0.44	200	200	29.1	29.1	223
SPT-11	49.9	55.5	12.8	12.8	0.44	0.44	200	200	29.1	29.1	251
SPT-12	56.6	42.3	12.8	12.8	0.43	0.44	200	200	29.8	29.1	204
SPT-13	45.1	46.7	12.8	12.8	0.42	0.43	200	200	30.5	29.8	224
SPT-14	46.1	45.7	12.8	12.8	0.43	0.43	200	200	29.8	29.8	228
SPT-15	50.5	42.8	12.8	12.8	0.43	0.43	200	200	29.8	29.8	251
SPT-16	48.1	37.7	12.8	12.8	0.43	0.42	200	200	29.8	30.5	208
SPT-17	46.3	41.2	12.8	12.8	0.42	0.43	200	200	30.5	29.8	201
SPT-18	50.2	36.7	12.8	12.8	0.43	0.43	200	200	29.8	29.8	223
SPT-19	58.4	42.5	12.8	12.8	0.44	0.43	200	200	29.1	29.8	194
SPT-20	56.5	40.5	12.8	12.8	0.44	0.43	200	200	29.1	29.8	220
SPT-21	56.2	35.6	12.8	12.8	0.42	0.42	200	200	30.5	30.5	208
SPT-22	52.2	40.6	12.8	12.8	0.43	0.43	200	200	29.8	29.8	199
SPT-23	52.0	38.5	12.8	12.8	0.43	0.43	200	200	29.8	29.8	220

Notes:

(1) Samples were all analyzed in duplicate: "Dup 1" indicates the first analytical measurement performed on the sample and "Dup 2" indicates the second analytical measurement performed on the same sample. Because of the large number of samples, the analytical determinations were performed in two separate batches. During the analytical determinations of the first and second batch, the results indicated by "Dup 1" and "Dup 2" respectively were obtained. In order to keep the samples fresh, the samples were frozen before the analysis of the second batch.

(2) This is the ratio of the Brix content and the Malic Acid concentration.

(3) Rounded to one decimal place in this table by the organizers of the SPT, in order to make the table more easily readable.

Appendix G: Peer-reviewed paper published from this research[1].
http://dx.doi.org/10.1039/C0AN00705F.

Analyst

Dynamic Article Links

Cite this: *Analyst*, 2011, **136**, 1313

www.rsc.org/analyst

PAPER

Improved evaluation of measurement uncertainty from sampling by inclusion of between-sampler bias using sampling proficiency testing

Michael H. Ramsey,*[a] Bastiaan Geelhoed,[a] Roger Wood[b] and Andrew P. Damant[b]

Received 10th September 2010, Accepted 11th January 2011
DOI: 10.1039/c0an00705f

A realistic estimate of the uncertainty of a measurement result is essential for its reliable interpretation. Recent methods for such estimation include the contribution to uncertainty from the sampling process, but they only include the random and not the systematic effects. Sampling Proficiency Tests (SPTs) have been used previously to assess the performance of samplers, but the results can also be used to evaluate measurement uncertainty, including the systematic effects. A new SPT conducted on the determination of moisture in fresh butter is used to exemplify how SPT results can be used not only to score samplers but also to estimate uncertainty. The comparison between uncertainty evaluated within- and between-samplers is used to demonstrate that sampling bias is causing the estimates of expanded relative uncertainty to rise by over a factor of two (from 0.39% to 0.87%) in this case. General criteria are given for the experimental design and the sampling target that are required to apply this approach to measurements on any material.

Introduction

There is now a general acceptance that the uncertainty of a chemical measurement is a crucial parameter for enabling the reliable interpretation of a measurement result for any purpose. It is also widely agreed that the measurement process begins at the time that the primary sample is taken, rather than when a derived laboratory sample arrives at the analytical laboratory.[1] It follows that the uncertainty of a measurement value has contributions from all of the steps within the analytical process. These steps will therefore include ones that are often excluded from that evaluation,[2] such as the primary sampling from a sampling target (*e.g.* a batch of food or an area of land) and the physical preparation of the primary sample (*e.g.* drying, grinding or splitting). Recent European Guidance[3] explains how this total uncertainty of measurement, including that arising from sampling, can be evaluated utilizing several published methodologies. The two main types of approach employed are either empirical[4,5] or based upon modelling.[6–8] The main empirical method that has been adopted is the 'duplicate method', in which one sampler takes duplicated samples from 10%, but no less than 8, of a selection of sampling targets. The limitation of this method, and nearly all of the other methods, is that they exclude the contribution to the uncertainty that arises from systematic effects in the sampling, quantified as sampling bias. However, the guidance[3] identifies that this limitation can be overcome by the use of multiple samplers, rather than a single sampler, in an experimental design equivalent to sampling proficiency tests (SPTs), but does not give a worked example. The main objective of SPTs has been to monitor, and enable improvements in, the quality of the processes of sampling and sample preparation that usually fall outside traditional analytical quality control procedures. Previous SPTs have been applied to a limited range of media (*e.g.* soil, water and gas) and their results have not yet been widely employed for the additional objective of uncertainty evaluation.

This paper aims therefore to:-

1. Describe how sampling proficiency test data can be used to make estimates of measurement uncertainty that include a contribution from sampling bias.
2. Identify generic design criteria for SPT that need to be met to enable both the evaluation of uncertainty and the effective scoring of the samplers, by consideration of published SPT studies.
3. Report a new SPT, that has been undertaken in the food sector, to exemplify the first objective.
4. Assess the usefulness and practicality of SPTs, and their potential role in both the evaluation of measurement uncertainty and in the assessment and improvement of the quality of primary sampling and sample preparation.

Previous sampling proficiency tests and their potential for evaluation of uncertainty

The concept of the sampling proficiency test (SPT) was proposed[9,10] as the sampling equivalent of the analytical proficiency test (APT). The APT has proved very effective in

[a]School of Life Sciences, University of Sussex, Falmer, Brighton, BN7 9QJ, UK. E-mail: m.h.ramsey@sussex.ac.uk
[b]Food Standards Agency, Aviation House, 125 Kingsway, London, WC2B 6NH, UK

improving and demonstrating the between-laboratory reproducibility of analytical measurements,[11] and the SPT was conceived to enable a similar improvement for sampling and physical sample preparation. In an APT, multiple portions of one test material are circulated between different analytical laboratories, which analyse the material for agreed analytes by any method that they consider appropriate. This test material is usually very stable and very homogeneous (*e.g.* finely ground if a solid), in order to get sufficient homogeneity within and between the multiple test portions. The APT is therefore very good at assessing the performance of analytical methods between different laboratories, but not their performance in the steps of primary sampling or physical preparation of the samples. It was envisaged that the SPT would fulfill that wider requirement. In the SPT the true value of analyte concentration to be estimated is that in the sampling target (*e.g.* a primary batch of material), rather than that in the laboratory sample or test material used in the APT. In metrological terminology this true value is effectively equivalent to the measurand, which is defined as the quantity intended to be measured.[12]

The first reported realisation of an SPT was on contaminated land,[13] which showed the feasibility of the concept and was used to generate scores for the samplers, but the results were not then used to estimate uncertainty. Several other SPTs have been reported since. One SPT used a synthetic reference sampling target[14] in which a known mass of analyte (barium) was placed in a known volume at a specific location within an area of land, to provide an estimate of the true or accepted value of the analyte concentration and also of its spatial position.[15]

The evaluation of the analytical component of measurement uncertainty using data from APTs is well established,[16,17] and it is based upon the Type A approach to the evaluation of measurement uncertainty.[18] This approach assumes that all known systematic effects have been corrected for, but one important issue is to develop methods for including unsuspected systematic effects (*e.g.* between-lab analytical bias) in the estimates of the uncertainty. One general approach has been to use a mathematical model to include bias. An example of this approach,[16] based upon an earlier study,[19] combines the bias using the equation:-

$$U = ku = k(u(R_w)^2 + u(\text{bias})^2)^{0.5} \quad (1)$$

where U is the expanded uncertainty for a coverage factor k, u is the standard uncertainty, and $u(R_w)$ is the within-laboratory reproducibility standard deviation. The uncertainty component for the bias u(bias) is given by

$$u(\text{bias}) = (\text{RMS}_{\text{bias}}{}^2 + u(C_{\text{ref}})^2)^{0.5} \quad (2)$$

where RMS_{bias} is the root mean square of the bias and $u(C_{\text{ref}})$ is the uncertainty component from the certified or nominal value of the APT test material. This approach recommends estimating the standard deviation of the bias using results from one lab over at least 6 rounds of an APT.

An alternative approach is to average the bias over the different PT participants and include that as a 'typical' value of the bias for an application of this particular method. This is the approach that was used in the first estimate of uncertainty made from SPT results.[20] However, instead of adding the bias using a model, it was added empirically using the between-participant variance from the previously published SPT measurement results, in this case for contaminated land. Robust analysis of variance (ANOVA) was used to quantify the various components of the uncertainty in a way that down-weights the effect of results from any outlying participants (as described below). A subsequent SPT also combined systematic errors in this way, but in addition it combined results from four different analytes to increase the degrees of freedom.[21]

SPTs have also been conducted on the measurement of contaminants in other media, such as landfill gas,[22] and canal water[23] that have temporal rather than spatial variability is the main source of uncertainty. This demonstrated that this approach can be adapted to these substantially different materials, and in one case, used for the estimation of uncertainty in a temporally variable system.[22] The wider applicability, feasibility and usefulness of SPTs in general will be assessed using the results of these previous studies together with the first reported SPT in the food sector on the determination of the moisture content in fresh wholesale butter, reported below.

Methods

Criteria for design of an effective SPT

The criteria for the design of an effective SPT (Table 1) have been derived from consideration of the findings in the previous SPTs and from statistical considerations. The degree to which these selection criteria can be met will vary between different types of targets and in different sectors of applications. Not every criterion has necessarily to be met in full, so the degree to which they need to be met will be discussed.

One example of the use of statistical considerations is the requirement that there be at least 8 samplers in an SPT (criterion 1), which is based upon the consideration of achieving a sufficiently

Table 1 Criteria for an effective SPT

#	Criteria
1	Size of the target must be large enough to allow sampling by at least 8 samplers, without significantly affecting the target.
2	The presence of an analyte that can be determined with lower enough analytical uncertainty to quantify sampling uncertainty.
3	Sufficient temporal stability of the sampling target (or else special procedures adopted to compensate for this[22]).
4	A level of heterogeneity of the analyte in the sampling target that is detectable with the proposed method of chemical analysis.
5	The presence of a recommended sampling protocol applicable to the selected analyte.
6	Availability of appropriate samplers with the required technical skills.
7	The potential usefulness for improving performance of the samplers, and accessibility of the sampling target.
8	Required time to perform the sampling protocol and duration of the SPT.
9	Acceptable level of cost due to damage to the target material and its packaging, during the SPT.
10	Independence of the samplers participating in the SPT.
11	Agreement of the owner of the sampling target and any sponsor of the SPT.
12	Independence of the different contributors to the SPT.

small confidence interval on the variance estimates.[24] Temporal stability of the sampling target is very important (criterion 3), especially where spatial variability is being characterised.[15] However some systems, such as landfill gas, are naturally temporally variable and the design of the SPT can be modified to make allowance for this criteria not being met.[22] The ideal situation is still, however, that the composition of the target is kept constant over the duration of the SPT.

In most cases, a perfectly homogeneous sampling target will not allow an SPT to show differences between the performance of the samplers. One exception is bias caused by variable contamination of the samples, but bias from most other causes will not be detected, so sufficient heterogeneity is required to cover all possibilities. A minimum amount of heterogeneity of the analyte distribution in the target (criterion 4) is needed in a sampling target, therefore, for the different behaviour of the samplers to be quantified. The following study will demonstrate that the absolute value of the heterogeneity can be relatively low (*e.g.* 0.08% RSD), especially if the analytical method is very precise (*e.g.* 0.17% RSD). This reflects therefore the related requirement for the presence of an analyte that can be determined in the sample (criterion 2), with an acceptable detection limit (*i.e.* with an analytical uncertainty that enables the quantification of the between-sampler uncertainty).

The presence of a recommended sampling protocol applicable to the selected analyte (criterion 5) enables samplers to follow pre-designed written instructions. However, cases have been known where only verbal protocols were available, and the SPT was then able to investigate the effect of a lack of written instructions. The availability of appropriate samplers (criterion 6) refers not just to their number (at least 8, as in criterion 1) but also to their level of technical skill and their ability to participate at the required time and place. It is desirable that the samplers be properly trained, but assessing the effectiveness of any training, especially if it is of short duration, can be a suitable objective for an SPT. A similar factor is to aspire to independence between the samplers (criterion 10). This applies during the SPT, when care needs to be taken so that no sampler can observe how another sampler is taking their samples. It also applies to the previous training, where ideally they should not have been recently trained by the same person. The effects of joint training of the samplers could be investigated using and SPT, but the resultant estimates of uncertainty from this sampling protocol may be expected to be lower than that when independent samplers are operating.

The potential usefulness for improving performance of the samplers (criterion 7) refers to the degree of flexibility in the system. For example, all of the human samplers might use one mechanical sampling device, which may be badly designed and located. In this case the measurements of all of the samplers in an SPT may agree well, but they may all be unrepresentative of the bulk composition of the sampling target. An example of this was a 20 000 L vertical tank of unstirred cloudy fruit juice on which the single sampling tap was located one third of the way up one side of the tank, which potentially under-represented the suspended matter content.[25] The individual human samplers might not therefore be in a position to improve their technique as a result of such an SPT. However the SPT organisers can help identify ways in which the overall quality of the sampling could be improved by installing a better designed mechanical sampling device, or choose a different target for the intended SPT. Closely related to this is the accessibility of a sampling target, which can be compromised when some of the batch cannot be accessed by the sampler. A similar example would be where the target material is kept in multiple sacks in a large pile, but the samplers can only access the sacks on the outside of the pile. As in the case above, this is an example of a poor sampling protocol that will affect the results of routine sampling as well as the SPT. The situation needs to be commented upon by the SPT organisers, and a different target chosen.

The time required to perform the sampling protocol (criterion 8) needs to not be excessive, so that the material does not change significantly in composition (criterion 3), and the cost to the SPT does not be thereby become unacceptable. This is closely related to the duration of the SPT, but obviously also depends on the number of samplers. There is also usually some level of damage to the target material, or its packaging, during the SPT (criterion 9). The SPT can be designed to reduce such damage as far as possible, whilst not deviating appreciably from the routine sampling protocol. The costs arising from the loss of target material during the SPT can be thereby minimized and included in the cost of running the SPT.

Clearly it is essential to secure the agreement of the owner of the sampling target and any sponsor of the SPT (criterion 11). Practical experience has shown that the independence of the different contributors to the SPT (criterion 12) is also essential. For example, if the organisation that performs the chemical analysis is also a participant in the SPT, or the manufacturer or owner of the sampling target, then there may be a conflict of interest in reporting measurement that affects the commercial interest of that organisation.[25] It is desirable, therefore, that all of these roles are kept in different organisations.

General experimental design of SPTs

The general experimental design used for most SPTs is a balanced design in which each sampler takes two samples, by independent interpretation of the protocol, both of which is analysed twice for each target analyte (Fig. 1). The analytical duplicates enable a separation of the contribution to the uncertainty from the analytical repeatability. This is required to ensure that repeatability is small enough to enable the quantification of the sampling variances. The duplicated sampling by each sampler provides an estimate of the within-sampler variance. The within-sampler variance is affected by the heterogeneity of analyte concentration within the sampling target. This heterogeneity is a key difference between and an SPT and an APT because in an APT heterogeneity is minimised as far as possible, but in an SPT it is an essential requirement (criterion 4 in Table 1). The experimental design has therefore to enable this heterogeneity to be quantified and separated so as not to affect the samplers' scores.

The between-sampler variance reflects the extra factors such as the between-sampler bias that can arise from systematic effects in either the sampling or the chemical analysis, depending on how the SPT is designed. The chemical analysis can either be undertaken by each of the participants separately, together with a matched reference material,[13] or the samples from all of the participants can be collected centrally and chemically analysed in

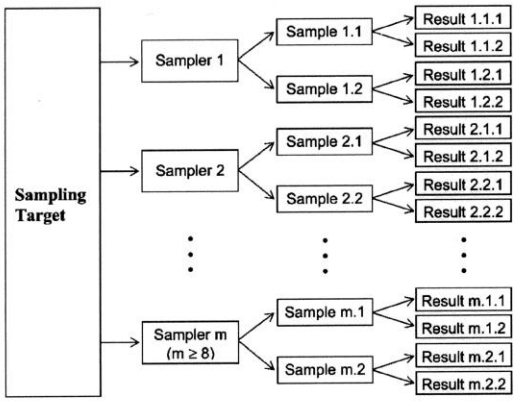

Fig. 1 Experimental design typically applied in sampling proficiency tests (SPTs), adapted from[35].

one laboratory under repeatability conditions.[15] The former approach could be called more precisely a 'measurement proficiency test' (MPT), as each participant undertakes the whole of the measurement process. The latter approach eliminates the effect of any analytical bias between the participants, helping the performance score to reflect only the participant's sampling performance. However, the centralised analysis in the latter approach will also tend to reduce the estimate of the total measurement uncertainty to an unrealistic level, unless it is dominated by the sampling component, or the analytical component is added independently.

Specific experimental design of SPT on moisture in butter

The determination of moisture in butter is an important parameter considered by the UK Rural Payments Agency for the acceptability of butter, either for the intervention scheme[26] or as 'butter for manufacture'.[27] A study of previously frozen butter[28] showed that the measurement uncertainty from sampling was low for moisture (U_{samp} = 2.5% relative or 0.39% m/m), due primarily to low levels of heterogeneity in the spatial distribution of the moisture in the butter. However, the closeness of the measured concentration (15.76% m/m in this case) to the legal limit (16% m/m) makes the sampling an important source of uncertainty. This is because it contributes 96% of the measurement uncertainty, and can therefore lead to misclassification of the butter against the legal limit.

The sampling target selected for this SPT was broadly defined as a batch of 20.1 t of fresh butter made in a creamery in SW England. The experimental design used is essentially the general one described in Fig. 1, with 9 participant samplers (identified by letters A–I), and full details are reported elsewhere.[25] The butter is packaged in the normal way, in around 804 cardboard boxes each containing a 25 kg block wrapped in a blue polythene sheet, and stored in a cold room chilled to 0–5 °C, on separate pallets to allow access to all boxes. This SPT design generally meets all of the selection criteria (Table 1), but with two partial exceptions. The 100 g samples were taken with a metal 'trier', that was a half-round tube around 30 cm long, with a 2 cm internal diameter and a handle at one end. The trier was inserted into the selected 25 kg block of butter at an angle of around 45° (variable between samplers 0–90° observed), twisted and then removed. The top and bottom portions of the resultant core (~20 cm, but variable) were rejected and the rest of the core transferred into a plastic pot (volume around 500 mL). Two cores were taken from each 25 kg block, to give a sample mass of around 100 g in each pot. Each sampler sampled six blocks, randomly selected from the whole batch, that generated a sample mass of around 600 g. This composite sample represented the 20.1 t of product, giving a sample ratio of 0.003% m/m (600 g/20 100 000 g). Criterion 1 was therefore satisfied. Each sampler took around 75 min to execute the sampling protocol, and the whole SPT took 5 hours to complete (criterion 8), during which time there was no detectable change in the concentration of the analyte (criterion 3). The variability in the position and angle at which the trier was inserted, and in the length of the discarded portion of the core, left room for improving the consistency of the sampler's performance (criterion 7).

The two criteria that were not fully met included the sufficient heterogeneity of the material (criterion 4). The heterogeneity at around 0.08% RSD was low as expected, but the precision of the gravimetric method (0.17% RSD) was not low enough to quantify this heterogeneity accurately, and should ideally have been less than 10% of that value (*i.e.* 0.008% RSD). The analytical precision was low enough to quantify the between-sampler effect (0.39% RSD, 0.78% at 95% confidence). This made the presence of non-proficient samplers or bias between samplers detectable, but potentially hard to quantify accurately (discussed below). The independence of the samplers (criterion 10) was the second criterion that was not fully met. Although the samplers were mainly from different local authorities, they had received some degree of common training, causing a possible lack of independence. This possibility was supported by the observation that many of the samplers adopted a very similar protocol that was implemented in a very similar way, despite the ambiguity in the written version of the protocol.[29] The importance of the selection of participant samplers with 'required skill' (criterion 6) was investigated by including one non-professional and relatively untrained sampler (identified by letter I). This sampler was included to see if the SPT could detect this nominal lack of sampling proficiency.

The routine sampling protocol normally employed in the factory was slightly modified for all of the samplers, to enable the more accurate evaluation of the uncertainty and the scoring of the participants. Each sampler took two sets of six increments from the whole batch, the second set constituting the duplicate sample. This was different from the routine protocol in which one set of six increments was taken, and then divided into two sets of three increments that were combined physically to represent the two halves of the whole 20 t target. Routinely, the position of the split between the two halves of the batch was made at a random point based upon the location of the third and fourth increments. This latter routine practice caused some ambiguity in the definition of the sampling target, which was resolved in the data processing, as explained below. This illustrates that although methods for the evaluation of uncertainty can generally be applied without changing the essence of the sampling protocol, there are some situations where the routine

protocol has to be slightly modified to enable the methods to be applied rigorously.

The primary samples of butter from all of the 9 samplers were collected and the moisture content determined gravimetrically[30] in one laboratory, under randomized repeatability conditions. The moisture concentration values (expressed as unrounded mass fraction %m/m) were pre-processed to conform to a sampling target consisting of the full 20.1 t. The moisture content of first 6 increments taken by each sampler was averaged to represent the full batch (m.1.1 in Fig. 1), and similarly their analytical duplicates (m.1.2). The second lot of 6 increments, that had been taken across the batch as the sample duplicate, were also averaged (m.2.1), and their analytical duplicates (m.2.2), to give the four measurements for each participant sampler.

Scoring of the SPT

The performance z-score (z) for each participant is calculated using an equation derived from that generally used for analytical proficiency testing, as defined by the International Harmonized Protocol,[11] and discussed in more detail elsewhere[25]

$$z = (x - X)/\sigma_{\text{target}} \quad (3)$$

where x is the participant's measurement result, X is the assigned value and σ_{target} is the target value for the standard deviation. In this case the assigned value is based on a consensus value expressed as the robust mean[4] of all of the measurements from all of the participants.

The target standard deviation (σ_{target}) is derived from the fitness-for-purpose standard deviation specified by the butter manufacturer (0.07%). This figure was adjusted to account for the modification made to the routine sampling protocol in this experimental design. Because six samples are composited (numerically), rather than the normal three (physically), a division by $\sqrt{2}$ was used to approximate a new value for the fitness-for-purpose standard deviation for a sample with six increments as $\sigma_{\text{target}} = 0.07/\sqrt{2} = 0.049\%$. This factor was derived from sampling theory, in which the variance of a measurement made on a sample is, under certain conditions, inversely proportional to the mass of the sample.[31,32]

Evaluation of measurement uncertainty

The evaluation of the measurement uncertainty using the data from this SPT is based on the output from the robust analysis of variance (RANOVA). Robust estimators are used to accommodate the effects of up to 10% of outlying values that do not conform to a Gaussian distribution.[4,33] The total variance is resolved into 3 components: analytical, within-sampler and between-sampler,

$$\sigma_{\text{total}}^2 = \sigma_{\text{between-sampler}}^2 + \sigma_{\text{within-sampler}}^2 + \sigma_{\text{analytical}}^2 \quad (4)$$

Estimates of the variance of the population (σ) made experimentally (s) give the working relationship:

$$s_{\text{total}}^2 = s_{\text{between-sampler}}^2 + s_{\text{within-sampler}}^2 + s_{\text{analytical}}^2 \quad (5)$$

The analytical variance is based upon the difference between the analytical duplicates (*i.e.* m.1.1 and m.1.2, and also m.2.1 and m.2.2, for sampler m, in Fig. 1). It gives an estimate of the portion of the measurement uncertainty arising from the repeatability of the analytical method, but ignores the contribution from any analytical bias. The within-sampler variance comes from the duplicated samples taken by each sampler (*i.e.* m.1 and m.2 in Fig. 1) after subtraction of the analytical variance. This is equivalent to the uncertainty estimate that would have been expected if the 'duplicate method' was applied to this one target using a single sampler,[3] rather than to the minimum of 8 targets usually recommended. The ANOVA separates the between-sampler variance as that which remains when the within-sampler and analytical variances are subtracted from the total variance. This between-sampler variance quantifies the contribution to the uncertainty caused by factors such as the sampling bias between the participants. The contribution from any analytical bias between the participants has been removed in this case, by the experimental design centralizing the analysis in one laboratory in a randomised batch under repeatability conditions.

The variance components from eqn (5) are quantified with RANOVA, and converted into the estimates of the corresponding measurement uncertainties. For this purpose the estimate of analytical repeatability, estimated across all of the participants' test materials, is included as part of the 'within-sampler' uncertainty.

$$U_{\text{within-sampler}} = 2\sqrt{(s_{\text{within-sampler}}^2 + s_{\text{analytical}}^2)} \quad (6)$$

To estimate the multi-sampler measurement uncertainty ($U_{\text{multi-sampler}}$) and standard deviation ($s_{\text{multi-sampler}}$), the between-sampler estimate ($s_{\text{between-sampler}}$) is added:

$$U_{\text{multi-sampler}} = 2\sqrt{(s_{\text{between-sampler}}^2 + s_{\text{within-sampler}}^2 + s_{\text{analytical}}^2)} \quad (7)$$

and

$$U_{\text{between-sampler}} = 2\sqrt{(s_{\text{between-sampler}}^2)} \quad (8)$$

Results and discussion

The four raw measurement results for the gravimetric determination of the moisture mass fraction were calculated for the arithmetic composite samples (Table 2). They show good analytical repeatability, quantified by the RANOVA as 0.35% relative, at 95% confidence. This is low enough to quantify the between-sampler effect (0.78% relative), but not the heterogeneity estimated from the within-sampler effect (0.17% relative, both at 95% confidence).

SPT scoring

In this design of SPT, each participant produces four z-score results based upon the duplicated analytical measurements on both primary samples (Fig. 1). One single z-score value for both samplers I and G are numerically greater than 2, suggesting the possibility of bias (Table 3). A more robust approach uses an

Table 2 Concentrations of moisture in butter (expressed as a mass percentage, %m/m), for the Sampling Proficiency Test, expressed for the single sampling target of 20.1 t. The measurements are quoted in the unrounded form reported by the laboratory to enable the consequent calculations to be reproduced if necessary

Sampler (m)	Analysis result m.1.1	Analysis result m.1.2	Analysis result m.2.1	Analysis result m.2.2
A	15.4741	15.4155	15.4972	15.4796
B	15.3655	15.3257	15.3653	15.3373
C	15.4417	15.4069	15.4552	15.4518
D	15.4161	15.4134	15.4486	15.4143
E	15.4085	15.3675	15.4392	15.4060
F	15.4148	15.3876	15.4176	15.3473
G	15.4959	15.4757	15.4853	15.5185
H	15.3673	15.3732	15.3720	15.3427
I	15.3214	15.2779	15.3424	15.3721

Table 3 Scores for SPT on moisture in butter SPT. The individual *z*-scores are calculated for both measurements (m.1.1 and m.1.2) for the two samples (m.1 and m.2) taken by each sampler (m) (Fig. 1). Individual |*z*-score| > 2 (shown in italics) suggests that samplers G and I may be non-proficient, but in opposite directions, in terms of their bias. This conclusion is confirmed more strongly by the rescaled sum of *z*-scores (RSZ), where the higher scores (|*z*-score| > 3, shown in bold) reveal persistent bias across the four results for each participant

Sampler	First *z*-score m.1.1	Second *z*-score m.1.2	Third *z*-score m.2.1	Fourth *z*-score m.2.2	Rescaled sum of *z*-scores
A	1.36	0.18	1.83	1.47	*2.42*
B	−0.83	−1.64	−0.84	−1.40	*−2.35*
C	0.71	0.00	0.98	0.91	1.30
D	0.19	0.14	0.85	0.15	0.66
E	0.04	−0.79	0.66	−0.01	−0.06
F	0.16	−0.39	0.22	−1.20	−0.60
G	1.80	1.39	1.59	*2.26*	**3.52**
H	−0.80	−0.68	−0.70	−1.29	−1.73
I	−1.72	*−2.60*	−1.30	−0.70	**−3.16**

overall score based upon all four *z*-scores per sampler (*i.e.* $m = 4$) by combining them into a single 'rescaled sum of *z*-scores' ($RSZ = \sum z/\sqrt{m}$) which has been recommended to detect small consistent bias in one system.[34] In this way, *z*-scores of opposite signs cancel out, but consistent bias is revealed. These RSZ scores (right hand column of Table 3) do show that samplers I (−3.16) and G (+3.52) have outlying scores, but with opposite signs to each other. This suggests that the non-professional sampler (I) is consistently under-reporting the moisture content, whereas one professional sampler, G, is consistently over reporting, compared with the consensus. Interpreting these RSZ scores as standard normal deviates, as suggested,[34] would indicate that these values show a degree of non-proficiency. This conclusion is, however, somewhat tentative, as there is a lack of independence between the four measurements for each participant, as two of the measurements arise from the same primary sample.

Uncertainty evaluation

Considering the estimates of expanded uncertainty from all nine SPT participants, the between-sampler uncertainty (0.12% m/m, Table 4) is twice as large as the within-sampler uncertainty (0.06% m/m). This means that the multi-sampler estimate of uncertainty from the SPT (0.13% m/m) is more that twice the size (2.2 times) of that estimated using a method equivalent to the duplicate method. This indicates that the value of the overall measurement uncertainty, estimated using the SPT approach, is substantially higher (by 120%) than that estimated using the equivalent of the accepted 'duplicate method' and is therefore more likely to be realistic because it includes the contributions from any sampling bias of the participants.

It is interesting to consider whether the estimates of uncertainty are being unduly affected by the non-proficient samplers (*i.e.* with |RSZ| score above 3). When the measurements for the non-professional sampler (I, with RSZ = −3.16) are removed, the multi-sampler, between-sampler and within-sampler estimates of uncertainty all fall by around 10%. However, the removal of both non-proficient samplers (*i.e.* I and G in Table 3) barely affects the within-sampler uncertainty (−5%), but the between- and multi-sampler uncertainties both fall by around 20%. These results broadly suggest that the bias of the non-proficient samplers was having an effect on the overall uncertainty estimate. The relative expanded uncertainty of the measurements for all samplers is estimated to be 0.87%, but the estimate for the seven proficient samplers only is 0.69%. One question is which of these estimates is more reliable for routine application of this measurement procedure. Perhaps 0.69% is most appropriate for samplers who are regularly found proficient in SPTs, but 0.87% may be a more reliable estimate for routine samplers who do not benefit from the feedback generated by participating in regular SPTs. Either of these multi-samplers values will be preferable to the over-optimistic within-sampler estimate of around 0.37% (*i.e.* 0.06% m/m). There is still an approximate factor of two (1.9) between the estimates of uncertainty using the multi- and within-sampler approaches of just the seven proficient samplers, so between-sampler effects such as sampling bias are not just caused by the outlying sampler with high |RCZ| scores.

It could be argued that if 8 different sampling targets had been used for the application of the duplicate method, as recommended,[3] then the uncertainty estimate by the 'duplicate approach' may have been higher than found here with just one target, especially if this particular batch was untypically homogeneous in its moisture content. Using 8 different targets might therefore be expected to make a smaller difference between the results by the two methods. However, the fact that the estimate uncertainty doubled on this one particular batch suggests that this two-fold difference is due to factors such as the presence of sampling bias, and not due to any potential abnormal heterogeneity of moisture in the batch. Ideally, 8 sequential SPTs could be conducted on 8 different sampling targets of nominally the same specification to resolve this question experimentally. It can also be argued that the presence of a non-professional sampler (participant I in Table 2) in this SPT made the results unreliable. However, in the real world there will always be a tendency for relatively inexperienced samplers, or ones trained in different ways, to take samples on some occasions. It is therefore interesting that the bias that such samplers may contribute to the estimate of uncertainty of 'typical' measurements, despite its down-weighting by the use of robust ANOVA. Moreover the |RSZ| score for one of the professional samplers (G) was also over three, suggesting that non-proficient scores are not limited to the non-professional sampler.

Incidentally, to make these estimates of uncertainty (Table 4) applicable to the current routine sampling procedure (*i.e.* taking 3 increments from ~10 t, rather than six increments from entire 20 t), it can be estimated that they would need to be multiplied by the square-root of two (1.414), using sampling theory as discussed above.

General discussion

The justification for using the total variance to derive the total measurement uncertainty (in eqn (4)) is that it includes all of the components including contributions from any sampling bias averaged across all of the participants. When any samplers are considered to be untypically biased (*e.g.* with |RSZ| > 3) then the uncertainty can be re-calculated with the measurements for those participants removed. In the resultant dataset, where none of the participants has any significant sampling bias, by this criterion, it might be expected that the estimate of multi-sampler uncertainty would not be significantly greater than that within-sampler, but this does not seem to be the case. This suggests that there may be low levels of sampling bias present even were the participant gets an 'acceptable' |RSZ| score.

It is possible that one participant could contend that they had really lower uncertainty than that represented by the multi-sampler (or even within-sampler) estimate. This contention might arise when they have a low |RSZ| score, perhaps in a number of sequential SPTs. In that case, the key issue would be to detect whether that participant has any undetected sampling bias. This would be difficult to detect, but one option could be for that participant to measure the target analyte on a suitable reference sampling target that has a certified value for that analyte concentration, if one is available. In that case using a reference sampling target for the detection of sampling bias is the sampling equivalent of using a reference material for the detection of analytical bias.[10]

Making general deductions from any single SPT is inevitably limited. For example the analytical contribution to the uncertainty is very low in this case, due to the high precision of the gravimetric determination of the analyte (0.35% at 95% confidence). In a different situation where the precision of the analytical method is much higher (*e.g.* 10%), but the sampling uncertainty is similarly low (*e.g.* <1%), then no information on the proficiency of the samplers would be detectable. This relates to the second selection criterion, where an analytical method with a lower precision will be required to enable such an SPT to give useful results. If this situation is viewed as a measurement proficiency test (MPT), however, then the results would be useful in showing participants that achieving a lower score would depend on reducing their analytical rather than their sampling uncertainty.

Further research could include a systematic comparison of the uncertainty values that are estimated using SPTs and those derived from the alternative model-based method of calculation (*e.g.* eqn (1)), when it is applied to the whole measurement system, including the sampling. Such a comparison would ideally include not just the resultant uncertainty values but also the confidence intervals of these estimates. Confidence intervals have been calculated for classical estimates of variance,[24] but not yet for robust estimates.

The cost of uncertainty evaluation is one factor that limits the evaluation and reporting of measurement uncertainty values. This was an issue for laboratories that aimed to estimate just the analytical component of the uncertainty, but it is even more relevant to the sampling process. The extra cost of implementing the duplicate method varies upon the size of the batch, ranging from 24% for a batch of 100 materials to over 300% for small batches of less than 8 materials. A simplistic calculation would suggest that the extra cost of implementing an SPT is over 700%, due to the need for a minimum of 8 samplers, and the requirement for the participants to travel to the sampling target, but this overlooks two factors. Firstly the duplicate method does also require at least 8 sets of replicates sampling and analysis, so the SPT is only different in that these are taken by different samplers, rather than the same one. Secondly, if SPTs are being run routinely for other purposes, such as monitoring performance (as is the case for APTs), then there is very little extra cost in calculating the uncertainty from the same measurements.[17]

The cost of evaluating uncertainty, by whatever method, can be justified by considering the financial consequences of either ignoring uncertainty altogether or leaving out one component, such as unsuspected sampling bias. These consequences might arise for actions taken unnecessarily or from mistaken inaction. It is for the situations where the financial consequences are very high that use of SPTs for uncertainty estimation will be more appropriate.

General usefulness and practicality of SPTs

Effective SPTs have previously been reported in several different media, ranging from the temporally stable and spatially heterogeneous (*e.g.* contaminated land) to the temporally heterogeneous (*e.g.* landfill gas and water). The food SPT described here demonstrates that even an almost homogeneous material (U_{samp} < 1%) can give useful information about sampling proficiency, if the analytical method is precise enough. It has been shown that in

Table 4 Estimates of uncertainty for the measurement of moisture in butter, subdivided into within-sampler, between-sampler and the overall multi-sampler. Data from Table 2 were used to calculate values for all 9 participant samplers and then with two participants with high RSZ-scores (>3 for I & G, in Table 3) removed progressively. The relative uncertainty is here expressed as a percentage of the robust average of the moisture content

	Absolute (%m/m)	Relative (%)	Absolute (%m/m)	Relative (%)	Absolute (%m/m)	Relative (%)
Number of samplers (n)	9	9	8 (−I)	8 (−I)	7 (−I and G)	7 (−I and G)
$U_{\text{within-sampler}}$	0.06	0.39	0.05	0.35	0.06	0.37
$U_{\text{between-sampler}}$	0.12	0.78	0.11	0.71	0.09	0.58
$U_{\text{multi-sampler}}$	0.13	0.87	0.12	0.79	0.11	0.69

the current and in previously published studies, that the SPT is not only practically achievable but also useful in many situations, provided that certain criteria are met to the appropriate degree (Table 1). An SPT can detect non-proficient samplers in a quantitative way that was not possible by any other means. The improvement in data quality that has been achieved in analytical laboratories by using APTs can potentially be extended out to include the sampling environment. Furthermore the intermediate area of physical sample preparation, whether it occurs in the field or after delivery of the sample to the lab, can also be brought within the measurement process. The impact of sampling preparation on the measurement result, and its uncertainty, can also be quantified using SPTs for the first time.

By including the otherwise un-detected sampling bias in the measurement uncertainty, the SPT can provide a much more robust estimate of uncertainty than is provided by the more widely used 'duplicate method'. Although more expensive to implement, there are some circumstances where the inclusion of sampling bias in the uncertainty estimate is essential to enable reliable decisions based upon measurements to be more reliable. For example where a batch of material is worth hundreds of millions of pounds/euros/dollars, then deciding the fate of the material often (*e.g.* a precious metal ore body) depends on having a reliable estimate of the uncertainty of the analyte concentration.

Conclusions

Sampling proficiency tests (SPTs) can be used to make an estimate of the uncertainty of measurement that includes the contribution from sampling bias, which is excluded by most existing methods. Most current methods of uncertainty evaluation only include the contributions from random and systematic effects in chemical analysis (reflected in analytical precision and bias). The 'duplicate method' previously improved this by enabling the random effects from sampling (*i.e.* sampling precision) to be included into estimates of uncertainty, but specifically excluded sampling bias. Mathematical models have been proposed for including systematic effects, where the values are previously known, but SPTs give a direct estimate of uncertainty that includes sampling bias without prior estimation. The feasibility of SPTs has been extended to include an example from the food sector. A series of criteria has been identified that enable an effective SPT to be designed and implemented, and a suitable sampling target to be selected. The cost of estimating measurement uncertainty is recognised as an important issue, and it is evident that using SPT results is potentially a more expensive option than using the duplicate method, depending on the other purposes for which the SPT is being used. The duplicate method has one advantage in that it averages the uncertainty over eight different sampling targets. Experience with analytical PTs suggests that a better mode of application will ultimately be to use the results from a series of SPTs on several sampling targets to provide a more generally representative estimate of the uncertainty. However, the cost of this approach will be higher, so it will only be justifiable where the consequences of undetected sampling bias are significant in terms of altering the decisions made as a result of the measurements. The further benefit of SPTs is that the scores can be used to monitor and enable improvements in the quality of the sampling and sample preparation process, which usually falls outside the traditional measurement quality control procedures.

Acknowledgements

The authors wish to acknowledge the funding for this research from the UK Food Standards Agency (FSA, E01070) and valuable assistance from Ellis Evans (FSA), Roy Smyth (RPA), and Martyn Allcorn (Westbury Creamery). Mike Thompson is thanked for constructive discussions on calculating performance *z*-scores.

References

1 M. H. Ramsey, *Accredit. Qual. Assur.*, 2004, **9**, 11–12, 727–728.
2 *Eurachem/CITAC Guide: Quantifying Uncertainty in Analytical Measurement*, ed. S. L. R. Ellison, M. Roesslein and A. Williams, Eurachem secretariat, LGC Ltd, London, 2nd edition, 2000.
3 *Eurachem/EUROLAB/ CITAC/Nordtest/ AMC Guide: Measurement Uncertainty Arising from Sampling: A Guide to Methods and Approaches*, ed. M. H. Ramsey and S. L. R. Ellison, Eurachem, 2007, (http://www.eurachem.org/guides/UfS_2007.pdf).
4 M. H. Ramsey, *J. Anal. At. Spectrom.*, 1998, **13**, 97–104.
5 C. Gron, J. B. Hansen, B. Magnusson, A. Nordbotten, M. Krystell, K. J. Andersen and U. Lund *Uncertainty from SAMPLING–A Nordtest handbook for Sampling Planners on Sampling Quality Assurance and Uncertainty Estimation*, Nordtest TR604, Nordtest, Espoo, 2007.
6 P. De Zorzi, M. Belli, S. Barbizzi, S. Menegon and A. Deluisa, *Accredit. Qual. Assur.*, 2002, **7**, 182–188.
7 U. Kurfurst, A. Desaules, A. Rehnert and H. Muntau, *Accredit. Qual. Assur.*, 2004, **9**, 64–75.
8 P. Minkkinen, *Chemom. Intell. Lab. Syst.*, 2004, **74**, 85–94.
9 M. H. Ramsey, in *Sampling of Environmental Materials for Trace Analysis*, ed. B. Markert, VCH, Weinheim, 1994, pp. 93–108.
10 M. Thompson and M. H. Ramsey, *Analyst*, 1995, **120**, 261–270.
11 M. Thompson and R. Wood, *Pure Appl. Chem.*, 1993, **65**, 2123–2144.
12 *JCGM 200 International Vocabulary of Metrology–Basic and General Concepts and Associated Terms*, Joint Committee for Guides in Metrology, BIPM, Sèvres, 2008.
13 A. Argyraki, M. H. Ramsey and M. Thompson, *Analyst*, 1995, **120**, 2799–2804.
14 M. H. Ramsey, S. Squire and M. J. Gardner, *Analyst*, 1999, **124**(11), 1701–1706.
15 S. Squire, M. H. Ramsey, M. J. Gardner and D. Lister, *Analyst*, 2000, **125**, 2026–2031.
16 Eurolab, *Measurement Uncertainty Revisited: Alternative Approaches to Uncertainty Evaluation*, Technical Report 1/2007, Eurolab, Paris, 2007, pp.12.
17 J. Wallace, *J. Forensic Sci.*, 2010, **55**(3), 767–773.
18 *JCGM 100 Evaluation of Measurement Data–Guide to the Expression of Uncertainty in Measurement (GUM 2008)*, Joint Committee for Guides in Metrology, BIPM, Sèvres, 2008.
19 *Nordtest Handbook for Calculation of Measurement Uncertainty in Environmental Laboratories* Nordtest Technical Report 537, Nordtest, Espoo, 2007.
20 M. H. Ramsey and A. Argyraki, *Sci. Total Environ.*, 1997, **198**, 243–257.
21 P. Lischer, R. Dahinden and A. Desaules, *Sci. Total Environ.*, 2001, **264**, 119–126.
22 S. Squire and M. H. Ramsey, *J. Environ. Monit.*, 2001, **3**, 288–294.
23 M. N. Rietveld and A. Goerdajal, Laboratoriumevaluerend onderzoek, Project 476-Bemonsteren oppervlaktewater, Afdeling: WGML (Water en Gebruik, Monitoring en Laboratorium) Rijkswaterstaat Ministerie van Verkeer en Waterstaat. (2010 report of sampling proficiency test on composition of canal water—unpublished work, in Dutch).

24 J. A. Lyn, M. H. Ramsey, S. Coad, A. P. Damant, R. Wood and K. A. Boon, *Analyst*, 2007, **132**, 1147–1152.
25 B. Geelhoed and M. H. Ramsey, *Feasibility, Practicality and Usefulness of Sampling Proficiency Tests in the Food Sector*, Project Report E01070, Food Standards Agency, London.
26 Commission Regulation (EC) No 2771/1999, *Off. J. Eur. Union*, 24/12/1999, **L333**, 0011–0043.
27 Commission Regulation (EC) No 2571/97, *Off. J. Eur. Union*, 20/12/1997, **L350**, 0003–0035.
28 J. A. Lyn, M. H. Ramsey, A. Damant and R. Wood, *Analyst*, 2005, **130**, 1271–1279.
29 *Butter Sampling Code*, Rural Payment Agency Inspectorate, August 2003, p.15, unpublished.
30 BS5086: Part 2, *Analysis of Butter–Method for the Determination of Water, Solids-not-Fat and Fat Content (reference method)*, BSI, Bristol, 1984.
31 P. Gy, *Sampling of Particulate Materials–Theory and Practice*, Elsevier, Amsterdam, 1979, p.431.
32 J. Visman, *Mater. Res. Stand.*, 1969, **9**, 8.
33 Analytical Methods Committee, *Analyst*, 1989, **114**, 1699–1702.
34 Analytical Methods Committee, *Analyst*, 1992, **117**, 97–117.
35 M. H. Ramsey and M. Thompson, *Accredit. Qual. Assur.*, 2007, **12**, 503–513.

Appendix H: List of sample codes for second SPT on moisture in butter

sampler	duplicate number	increment	sample code	sampler	duplicate number	increment	sample code
A	1	1	SPT-34	E	2	1	SPT-93
A	1	2	SPT-14	E	2	2	SPT-70
A	1	3	SPT-1	E	2	3	SPT-107
A	1	4	SPT-77	E	2	4	SPT-20
A	1	5	SPT-61	E	2	5	SPT-17
A	1	6	SPT-75	E	2	6	SPT-45
A	2	1	SPT-46	F	1	1	SPT-85
A	2	2	SPT-30	F	1	2	SPT-99
A	2	3	SPT-101	F	1	3	SPT-95
A	2	4	SPT-21	F	1	4	SPT-15
A	2	5	SPT-38	F	1	5	SPT-50
A	2	6	SPT-100	F	1	6	SPT-83
B	1	1	SPT-90	F	2	1	SPT-11
B	1	2	SPT-56	F	2	2	SPT-5
B	1	3	SPT-106	F	2	3	SPT-84
B	1	4	SPT-92	F	2	4	SPT-10
B	1	5	SPT-6	F	2	5	SPT-44
B	1	6	SPT-37	F	2	6	SPT-59
B	2	1	SPT-94	G	1	1	SPT-108
B	2	2	SPT-57	G	1	2	SPT-88
B	2	3	SPT-40	G	1	3	SPT-3
B	2	4	SPT-48	G	1	4	SPT-33
B	2	5	SPT-8	G	1	5	SPT-78
B	2	6	SPT-76	G	1	6	SPT-18
C	1	1	SPT-87	G	2	1	SPT-52
C	1	2	SPT-26	G	2	2	SPT-27
C	1	3	SPT-42	G	2	3	SPT-104
C	1	4	SPT-79	G	2	4	SPT-19
C	1	5	SPT-31	G	2	5	SPT-65
C	1	6	SPT-102	G	2	6	SPT-43
C	2	1	SPT-60	H	1	1	SPT-51
C	2	2	SPT-55	H	1	2	SPT-49
C	2	3	SPT-36	H	1	3	SPT-73
C	2	4	SPT-62	H	1	4	SPT-82
C	2	5	SPT-58	H	1	5	SPT-91
C	2	6	SPT-69	H	1	6	SPT-25
D	1	1	SPT-53	H	2	1	SPT-41
D	1	2	SPT-105	H	2	2	SPT-39
D	1	3	SPT-72	H	2	3	SPT-22
D	1	4	SPT-64	H	2	4	SPT-32
D	1	5	SPT-80	H	2	5	SPT-68
D	1	6	SPT-98	H	2	6	SPT-96
D	2	1	SPT-71	I	1	1	SPT-74
D	2	2	SPT-9	I	1	2	SPT-16
D	2	3	SPT-67	I	1	3	SPT-54
D	2	4	SPT-103	I	1	4	SPT-2
D	2	5	SPT-23	I	1	5	SPT-4
D	2	6	SPT-89	I	1	6	SPT-86
E	1	1	SPT-66	I	2	1	SPT-7
E	1	2	SPT-24	I	2	2	SPT-13
E	1	3	SPT-12	I	2	3	SPT-47
E	1	4	SPT-81	I	2	4	SPT-63
E	1	5	SPT-28	I	2	5	SPT-29
E	1	6	SPT-35	I	2	6	SPT-97

Appendix I: Pre-prepared blank list for second SPT on moisture in butter

Sampler	SELECTED BOXES											
	BOXES THAT BELONG TO THE FIRST DUPLICATE SAMPLE:						BOXES THAT BELONG TO THE SECOND DUPLICATE SAMPLE:					
	Inc. 1	Inc. 2	Inc. 3	Inc. 4	Inc. 5	Inc. 6	Inc. 1	Inc. 2	Inc. 3	Inc. 4	Inc. 5	Inc. 6
A												
B												
C												
D												
E												
F												
G												
H												
I												

Appendix J: Timing of main activities during second SPT on butter

Time	Activity	Room
10-10.30	Briefing for 9 samplers, allocation of samplers i.d. letter A-I	Canteen
10-30-11.06	Selection of 6 sample boxes by each sampler (4 min each x 9) – writing box numbers on a pre-prepared blank list	Main Butter Store
11.06-11.36	Selection of duplicate 6 sample boxes by each sampler- writing box numbers on a second pre-prepared blank list	Main Butter Store
11.36 -12.36	Samplers:- extract their 12 selected boxes - Label boxes with their coloured labels (e.g. Green A1.1- A1.6, A2.1- A2.6, Blue B1.1- B1.6, B2.1- B2.6) - Put boxes onto their individual pallet (labelled A or B –H) Fork lift driver transports the 9 pallets to just outside sampling room	Main Butter Store
12.36 – 13.36	3 samplers at a time (working out of sight of each other) take samples from each of their first 6 (and then 2nd 6) boxes and place them in containers with pre-prepared labels (2 min per box). (Other samplers have lunch)	Sampling room
13.36 -14.36	2nd lot of 3 samplers take samples (as above). (Other samplers have lunch)	Sampling room
14.36 – 15.36	3rd lot of 3 samplers take samples (as above)	Sampling room
15.36 –16.06	Batch of all laboratory samples check off and dispatched to the laboratory used for this SPT (Eurofins)	Sampling room

Appendix K: Instructions for samplers for second SPT on moisture in butter

SAMPLING PROFICIENCY TEST AT THE CREAMERY - 8TH FEB 2007

1. Please select and take your samples as you would do normally.
2. Please select two sets of six boxes and write the box numbers for both sets on the pre-prepared list.
3. You will be given a colour pattern label you will label your selected boxes with to facilitate you finding your boxes.
4. After all participants have selected their boxes you will be asked to remove your selected boxes and put them on your pallet.
5. Your pallet will be transported to the sampling room where you will be asked to subsample each of your selected boxes in a separate area for each sampler.
6. If your box has also been selected by another sampler, please pass your box to that sampler after you have drawn all your subsamples from that box.
7. Please label your sample containers clearly and present them to the organizers of the Sampling Proficiency Test.

Appendix L: Analytical results for moisture for second SPT on moisture in butter

Sample ID[1]	Box Number[2]	Moisture 1[3], [4]	Moisture 2[3],[5]
A01	906	15.4439	15.4258
A02	983	15.2982	15.1956
A03	1100	15.5475	15.5086
A04	1193	15.5227	15.4893
A05	1294	15.6199	15.5698
A06	1579	15.4125	15.3038
A07	959	15.3678	15.4072
A08	1033	15.5078	15.5066
A09	1285	15.4785	15.4470
A10	1502	15.5617	15.4541
A11	1555	15.4731	15.4994
A12	1388	15.5942	15.5633
B01	1085	15.3268	15.3027
B02	948	15.3787	15.2804
B03	876	15.3128	15.3919
B04	1199	15.1783	15.1695
B05	1522	15.6387	15.4386
B06	1399	15.3577	15.3713
B07	1440	15.3475	15.3232
B08	1564	15.4263	15.4040
B09	1237	15.1704	15.2083
B10	841	15.2735	15.2559
B11	964	15.5486	15.4237
B12	1081	15.4256	15.4086
C01	1309	15.3049	15.2622
C02	1458	15.4622	15.4711
C03	819	15.5504	15.5044
C04	903	15.4670	15.3561
C05	968	15.3569	15.2908
C06	1033	15.5089	15.5567
C07	1116	15.3290	15.3333
C08	884	15.3983	15.4577
C09	1144	15.4510	15.4065
C10	1264	15.5361	15.5044
C11	1390	15.3707	15.4503
C12	1586	15.6458	15.5586
D01	1366	15.4007	15.3404
D02	1528	15.3932	15.4313
D03	1482	15.4159	15.4307
D04	1160	15.4175	15.4287
D05	925	15.4576	15.4739
D06	1080	15.4118	15.3756
D07	1401	15.5226	15.4164
D08	1519	15.4808	15.4196
D09	1281	15.4502	15.4353
D10	839	15.4889	15.4462
D11	959	15.3570	15.4439
D12	1040	15.3918	15.3241
E01	1442	15.4005	15.4049
E02	1400	15.2816	15.2441
E03	1320	15.5144	15.4260
E04	1240	15.3381	15.2051
E05	1199	15.5029	15.4457

E06	1159	15.4134	15.4794
E07	1119	15.4186	15.3769
E08	1079	15.4047	15.3751
E09	1039	15.5108	15.4605
E10	998	15.4645	15.4356
E11	958	15.3125	15.3282
E12	918	15.5239	15.4596
F01	1040	15.3189	15.2977
F02	908	15.2113	15.2455
F03	845	15.6027	15.5884
F04	1196	15.4849	15.4410
F05	1471	15.3363	15.2758
F06	1559	15.5347	15.4769
F07	1549	15.4524	15.3079
F08	1519	15.5401	15.4668
F09	1166	15.1962	15.1146
F10	879	15.4025	15.3920
F11	965	15.5297	15.4891
F12	1109	15.3845	15.3132
G01	1116	15.3675	15.4012
G02	1004	15.3401	15.3104
G03	884	15.5339	15.4764
G04	1200	15.5124	15.5034
G05	1326	15.6825	15.6887
G06	1448	15.5390	15.4743
G07	1397	15.5249	15.5877
G08	1519	15.5838	15.5968
G09	1157	15.3576	15.3891
G10	839	15.5391	15.5833
G11	959	15.4909	15.4799
G12	1079	15.4155	15.4744
H01	828	15.4130	15.4063
H02	908	15.3618	15.2902
H03	1106	15.0805	15.2153
H04	1221	15.4005	15.5444
H05	1357	15.4803	15.3461
H06	1579	15.4674	15.4368
H07	946	15.1140	15.1449
H08	1086	15.4365	15.4176
H09	1189	15.3834	15.3475
H10	1270	15.4630	15.4997
H11	1479	15.6674	15.5316
H12	1559	15.1674	15.1150
I01	1036	15.3635	15.3150
I02	912	15.3820	15.4635
I03	1160	15.3452	15.3219
I04	1471	15.1345	15.0026
I05	1180	15.4295	15.3333
I06	1321	15.2735	15.2309
I07	1116	15.4411	15.5306
I08	1401	15.4869	15.5518
I09	1522	15.3280	15.3887
I10	1146	15.2678	15.3039
I11	1260	15.1815	15.1959
I12	905	15.3493	15.2618

NOTES for Appendix L:

(1) The letter indicates the sampler: A=first sampler, B=second sampler, C=third sampler, D=fourth sampler, E=fifth sampler, F=sixth sampler, G=seventh sampler, H=eighth sampler, I=ninth sampler. The number represents the order in which the samples were taken, e.g. A10 represents the tenth increment drawn by sampler "A".
(2) Each box in the lot was numbered. The first box was numbered 805 and the last box was numbered 1608. Box numbers 806, 1007, 1208, and 1409 were absent.
(3) Units are in weight-percent (% m/m).
(4) This column gives the results for the first duplicate analysis on the sample.
(5) This column gives the results for the second duplicate analysis on the sample.

Appendix M: Results for SNF for second SPT on moisture in butter

Sample ID[1]	Box Number[2]	SNF 1[3], [4]	SNF 2[3], [5]
A01	906	1.8519	1.8801
A02	983	1.8941	1.7543
A03	1100	1.7107	1.6371
A04	1193	1.8288	1.8189
A05	1294	2.0214	2.0061
A06	1579	2.0328	2.0050
A07	959	1.8120	1.7983
A08	1033	1.7450	1.6769
A09	1285	1.8694	1.9503
A10	1502	1.6410	1.5910
A11	1555	1.9821	1.9403
A12	1388	1.9424	1.8477
B01	1085	1.7212	1.8239
B02	948	1.7435	1.7762
B03	876	1.8043	1.7448
B04	1199	1.9065	1.8793
B05	1522	1.7110	1.6590
B06	1399	1.8593	1.8308
B07	1440	1.9853	1.8440
B08	1564	1.8294	1.8052
B09	1237	1.9529	1.9575
B10	841	1.7518	1.7150
B11	964	1.6680	1.7150
B12	1081	1.9031	1.9000
C01	1309	1.9607	1.9080
C02	1458	1.6352	1.6043
C03	819	1.8634	1.8366
C04	903	1.8623	1.8893
C05	968	1.7841	1.7298
C06	1033	1.9984	1.9173
C07	1116	1.8248	1.7877
C08	884	1.7056	1.6693
C09	1144	1.7550	1.7440
C10	1264	1.8665	1.9309
C11	1390	1.7873	1.8575
C12	1586	1.7596	1.7571
D01	1366	1.7735	1.7699
D02	1528	1.9102	1.9076
D03	1482	1.8279	1.8184
D04	1160	1.8618	1.8402
D05	925	1.7694	1.8448
D06	1080	1.9239	1.9536
D07	1401	1.7844	1.7672
D08	1519	1.7710	1.7011
D09	1281	1.7916	1.8754
D10	839	1.9619	1.9296
D11	959	1.6732	1.6614
D12	1040	1.8359	1.8034
E01	1442	1.7116	1.7803
E02	1400	1.7654	1.7897
E03	1320	1.6247	1.6479
E04	1240	1.8453	1.8932
E05	1199	1.7671	1.7476

E06	1159	1.8780	1.8421
E07	1119	1.8210	1.8465
E08	1079	1.7500	1.7237
E09	1039	1.7954	1.7859
E10	998	1.7189	1.7774
E11	958	1.7367	1.8126
E12	918	1.7048	1.7334
F01	1040	1.8477	1.9042
F02	908	1.8576	1.9454
F03	845	1.6436	1.6990
F04	1196	1.7248	1.6956
F05	1471	1.8625	1.8402
F06	1559	1.7924	1.7422
F07	1549	1.6088	1.6685
F08	1519	1.7595	1.7225
F09	1166	1.6561	1.5994
F10	879	1.6733	1.6453
F11	965	1.7568	1.7396
F12	1109	1.8435	1.8748
G01	1116	1.7585	1.7782
G02	1004	1.8393	1.9101
G03	884	1.6211	1.7902
G04	1200	1.6734	1.6352
G05	1326	1.9329	1.8771
G06	1448	1.7476	1.7010
G07	1397	1.9295	1.9683
G08	1519	1.7870	1.7336
G09	1157	1.8874	1.9605
G10	839	1.7810	1.7756
G11	959	1.9413	1.9900
G12	1079	1.7950	1.8196
H01	828	1.9698	1.9536
H02	908	1.9006	1.9227
H03	1106	1.8698	1.9082
H04	1221	1.8232	1.7569
H05	1357	1.7120	1.9595
H06	1579	1.6296	1.6458
H07	946	1.8538	1.9838
H08	1086	1.7360	1.7700
H09	1189	1.7414	1.8567
H10	1270	1.7767	1.8315
H11	1479	1.8835	1.8219
H12	1559	1.8227	1.8138
I01	1036	1.8665	1.9785
I02	912	1.7751	1.8240
I03	1160	1.8090	1.8026
I04	1471	1.7885	1.6683
I05	1180	1.7415	1.6176
I06	1321	1.8766	1.8916
I07	1116	1.8048	1.7293
I08	1401	1.6934	1.6215
I09	1522	1.7554	1.7884
I10	1146	1.9144	1.9616
I11	1260	1.9227	1.9827
I12	905	1.9498	1.9921

NOTES for Appendix M:

(1) The letter indicates the sampler: A=first sampler, B=second sampler, C=third sampler, D=fourth sampler, E=fifth sampler, F=sixth sampler, G=seventh sampler, H=eighth sampler, I=ninth sampler. The number represents the order in which the samples were taken, e.g. A10 represents the tenth increment drawn by sampler "A".
(2) Each box in the lot was numbered. The first box was numbered 805 and the last box was numbered 1608. Box numbers 806, 1007, 1208, and 1409 were absent.
(3) Units are in weight-percent (% m/m).
(4) This column gives the results for the first duplicate analysis on the sample.
(5) This column gives the results for the second duplicate analysis on the sample.

References

[1] Ramsey M.H. Geelhoed B, Damant, A.P., Wood, R. (2011) Improved evaluation of measurement uncertainty from sampling by inclusion of between-sampler bias using sampling proficiency testing. Analyst, 136, 7, 1313 – 1321 http://dx.doi.org/10.1039/C0AN00705F

[2]Thompson, M and Wood, R (1993) "The international harmonized protocol for the proficiency testing of (chemical) analytical laboratories", Pure Appl. Chem., Vol 65, p. 2123-2144. (Also J. AOAC International 76, 926-940 (1993))

[3] Ellison S L R, Roesslein M, Williams A (eds) (2000) Eurachem/CITAC Guide: *Quantifying Uncertainty in Analytical Measurement*, Eurachem, 2nd edition, ISBN 0 948926 15 5. Available from the Eurachem secretariat, or LGC Ltd. (London).

[4] Ramsey MH (2004) When is sampling part of the measurement process? Accreditation and Quality Assurance: Journal for Quality, Comparability and Reliability in Chemical Measurement, 9, 11-12, 727 - 728

[5] Ramsey, M.H. (1994) 'Error estimation in environmental sampling and analysis' in Sampling of environmental materials for trace analysis. Markert B. (Editor) VCH, Weinheim, p 93-108.

[6] Thompson M. and Ramsey, M.H. (1995) Quality concepts and practices applied to sampling - an exploratory study, Analyst, 120, 261-270

[7] Argyraki, A, Ramsey, MH, Thompson, M (1995) Proficiency testing in sampling: pilot study on contaminated land, Analyst, 120, 2799-2803

[8] Argyraki, A and Ramsey, MH (1998) Evaluation of inter-organizational trials on contaminated land: comparison of two contrasting sites. In: Lerner, DN and Walton NRG (eds) 1998. Contaminated Land and Groundwater; Future Directions. Geological Society, London, Engineering geology Special Publications, 14, 119-125.

[9] Squire, S and Ramsey, M.H. (2001) Inter-organisational sampling trials for the uncertainty estimation of landfill gas measurements. Journal of Environmental Monitoring, 3, 3, 288 - 294

[10] Ramsey, M.H. and Argyraki A.(1997) Estimation of measurement uncertainty from field sampling: implications for the classification of contaminated land. Science of the Total Environment, 198, 243 – 257

[11] Squire, S., Ramsey, M.H., Gardner, M.J. and Lister, D. (2000) Sampling proficiency test for the estimation of uncertainty in the spatial delineation of contamination. Analyst, 125, Issue 11, 2026-2031

[12] Ramsey M.H. and Thompson M. (2007) Uncertainty from sampling, in the context of fitness for purpose *Accreditation and Quality Assurance: Journal for Quality, Comparability and Reliability in Chemical Measurement* 12, 503-513

[13] Thompson M., Willetts, P., Anderson S., Brereton P. and Wood R. (2002) Collaborative trials of the sampling of two foodstuffs, wheat and green coffee. Analyst, 127, 689-691

[14] Lyn, J.A., Ramsey, M.H., Coad, S., Damant, A.P., Wood, R. and Boon K.A. (2007) The duplicate method of uncertainty estimation: are eight targets enough? Analyst 132, 1147-1152 (DOI: 10.1039/b702691a)

[15] Ramsey M.H., and Ellison S. L. R.,(eds.) (2007) Eurachem/EUROLAB/ CITAC/Nordtest/ AMC Guide: *Measurement uncertainty arising from sampling: a guide to methods and approaches*, Eurachem, ISBN 978 0 948926 26 6. (http://www.eurachem.org/guides/UfS_2007.pdf)

[16] Ramsey, M.H., Lyn, J.A. and Wood, R. (2001) Optimised uncertainty at minimum overall cost to achieve fitness-for-purpose in food analysis. Analyst, 126, 10, 1777-1783

[17] Smith and Moss (1985) Mycotoxins: Formation, Analysis and Significance, Wiley, Chichester

[18] European Mycotoxin Awareness Network Website: http://www.mycotoxins.org/ Basic Factsheets/patulin (accessed 18/11/11)

[19] *Birkinshaw et al*. (1943) *Lancet* **245**, 625 cited in Steyn (Eds) (1980) The Biosynthesis Of Mycotoxins, Academic Press

[20] FDA Website: http://www.fda.gov/Food/FoodSafety/FoodContaminantsAdulteration/NaturalToxins/ucm212520.htm (site last accessed 16/11/11)

[21] COMMISSION REGULATION (EC) No 1425/2003 of 11 August 2003 amending Regulation (EC) No 466/2001 as regards patulin

[22] FSA Website http://www.food.gov.uk/multimedia/faq/patulin/ (site last accessed 16/11/11)

[23] COMMISSION DIRECTIVE 2003/78/EC of 11 August 2003 laying down the sampling methods and the methods of analysis for the official control of the levels of patulin in foodstuffs

[24] Analytical Methods Committee, *Analyst*, 1989, 114, 1699-1702.

[25] Ramsey, MH, Thompson, M, Hale, M (1992) Objective evaluation of precision requirements for geochemical analysis using robust analysis of variance, Journal of Geochemical Exploration, 44, 23-36.

[26] Hanson J.R.(2005) (Prof of Chemistry, University of Sussex) personal communication.

[27] Thompson, M, Ellison, SLR, Wood, R (2006) The international harmonized protocol for the proficiency testing of analytical chemistry laboratories, Pure Appl. Chem, 78, 145-196.

[28] Lyn J.A., Ramsey M.H., Damant A., and Wood R. (2005), Two stage application of the OU method: a practical assessment, *Analyst*, 130, 1271-1279

[29] Butter Sampling Code, Rural Payment Agency Inspectorate, August 2003, 15 pp.

[30] Analytical Methods Committee (1992) Proficiency testing of analytical laboratories: organization and statistical assessment, Analyst, 117, 97-117.

[31] Ruddle J (2008) Uncertainty from sampling and accreditation of samplers. Lecture given at Eurachem Workshop ' Sampling uncertainty and Uncertainty for compliance assessment' BAM Berlin, April 15-16, 2008

[32] ISO 11074-2: 1998 Soil Quality – Vocabulary. Part 2: Terms and definitions related to sampling, International Organization for Standardization, Geneva.

[33] ISO 3534-1: 1993 Statistics – Vocabulary and Symbols, International Organization for Standardization, Geneva
[34] AMC (2005) Analytical Methods Committee Technical Brief No 19. Terminology – the key to understanding analytical science. Part 2: Sampling and sample preparation (http://www.rsc.org/images/brief19_tcm18-25963.pdf).
[35] IUPAC (2005) Terminology in Soil Sampling (IUPAC Recommendations 2005), prepared for publication by De Zorzi P, Barbizzi S, Belli M, Ciceri G, Fajgelj A, Moore D, Sansone U, and Van der Perk M. *Pure and Applied Chemistry*, 77 (5), 827–841.