SAMPLING COLUMN

Sampling quality assessment: the replication experiment

Kim H. Esbensen[a] and Claas Wagner[b]
[a]KHE Consulting, www.kheconsult.com. E-mail: khe.consult@gmail.com
[b]Sampling Consultant. E-mail: cw@wagnerconsultants.com

This column gives an overview of an issue that has not received proper attention for decades, the issue of "replication". This issue turns out to be complex and there has been a lot of confusion in the literature. Three answers to what is often stated in response to the fundamental question: "what is replicated exactly?" are i) replicate samples, ii) replicate measurements or iii) replicate analysis (replicate analytical results). Upon reflection it is clear that these three answers are not identical. The often only implied understanding for all three cases is that a beneficial averaging is carried out with the connotation that important insight can be gained by "replication". By replicating the specific process behind replicated samples, measurements and results, some measure of variability is obtained; but a measure of what? There are many vague prerequisites and imprecise assumptions involved, which need careful analysis. For starters, i) addresses the **pre**-laboratory realm, while ii) and iii) play out their role **in** the analytical laboratory—but even here: are replicate analysis the same as replicate measurements?

Background

From the discipline of design of experiments (DOE) comes a strict conceptual understanding and terminology because of the controlled surrounding conditions. In the situation of chemical synthesis influenced by several experimental factors, temperature, pressure, concentration of co-factors for example, it is easy to understand what a replicate experiment means: one is to repeat the experimental run(s) under *identical* conditions for all controllable factors, taking care to randomise all other factors, in which case the variance of the repeated outcome, be it small or large, will furnish a measure of the "total experimental uncertainty", which will be larger than the strict analytical **repeatability**. In routine operations in the analytical laboratory, variability also reflects effects from other uncertainty contributions stemming, for example, from small-scale sampling of reactants involved, which may not necessarily represent completely "homogeneous stocks". Added uncertainty contributions may also occur from resetting the experimental setup—to what precision can one "reset" temperature, pressure, concentration levels of co-factor chemical species after having turned the setup off and cleaned all the experimental equipment? Still, such uncertainty contributions are usually considered acceptable parts of the total analytical error (TAE). Often all of the above turn out to be of small, or vanishing, effect because of the regular conditions surrounding a controlled DOE situation.

Stepping back one step, however, one might find it equally relevant to repeat the experiment by another technician, researcher and/or in another laboratory, enter the well-known analytical concept of **reproducibility**. There may be more, smaller or larger effects in this widened context, and careful empirical total effect estimations must always be carried out in order to arrive at a valid estimate of the augmented, effective TAE.

Behold the whole lot-to-analysis pathway

Below we address more external issues, not always on the traditional agenda for replication, in fact quite often left out, or forgotten.

There are in fact many scenarios that differ from a nicely bracketed DOE situation. Indeed most data sets do not originate exclusively from within the complacent four walls of an analytical laboratory. What will be described below constitutes the opposing end of a full spectrum of possibilities in which the researcher/data analyst must also recognise *significant* sampling, handling and other errors in addition to the effective TAE. The total sampling error (TSE) will include all sampling and mass-reduction error effects, all incurred *before* analysis. It is self-evident that these errors must also be included in realistic analytical error assessments; TAE alone will not give a relevant, valid estimate of the total effective effects influencing the analytical results. We are forced to be able to furnish a valid estimate of the total sampling-handling-analysis uncertainty estimate (GEE: =TSE+TAE).

The description below is supposed to deal comprehensively with the many different manifestations surrounding the replication issue, such that most realistic scenarios are covered. At the heart-of-the-matter is a key question: what is meant by "replicate samples"? This issue will appear more complex than may seem the case at first sight and will receive careful attention w.r.t. definitions and terminology. It will also transpire that this issue is intimately related to *validation* in data analysis, chemometrics and statistics.

Clarification

Upon reflection it will be appreciated that "replication" can concern the following alternatives in the lot-to-aliquot pathway from primary sampling to analytical result:

1. Replication of the primary sampling process
2. Replication starting with the secondary sampling stage (i.e. first mass reduction)
3. Replication starting with the tertiary sampling process (i.e. lab. mass reduction)
4. Replication starting with aliquot preparation (e.g. powder compactification)
5. Replication starting with aliquot instrument presentation (e.g. surface conditioning)
6. Replication of the analysis (measurement operation) only (TAE)

The last option is the situation corresponding to "replicate measurement" in the most restricted case. But does this mean that the analytical aliquot (the vial) stays in the analytical instrument all the time while the analyst simply "presses the button" say 10 times? Possibly; in which case this furthers a strict estimate of TAE *only*. However, it seems equally relevant to extract the vial and insert it in the instrument repeatedly, allowing a possible temperature variation to influence on TAE because this is a more *realistic* repetition of the general work and measurement process in any laboratory than simply leaving the test portion in the instrument. This is a first foray into what is known as "Taguchi thinking",[1] which opens up a focus on potentially influencing factors which are not embedded in the experimental design explicitly. Clearly this kind of external thinking is relevant in many situations and should therefore be included in the replication approach. One important dictum of Taguchi's is: do not necessarily look only for optimal results (which *may* have large variability), but to results where the response variability is low over a large span of the experimental domain (even if less optimal). This is a clever way of gaining more information about the process involved, be this a production or manufacturing process, or the analytical process itself. Certain scepticisms have been voiced regarding the merits of this approach, but we will let the reader decide on this matter.

Opening up for the relevance of this type of *perturbation* of the analytical process, to another analyst it may appear

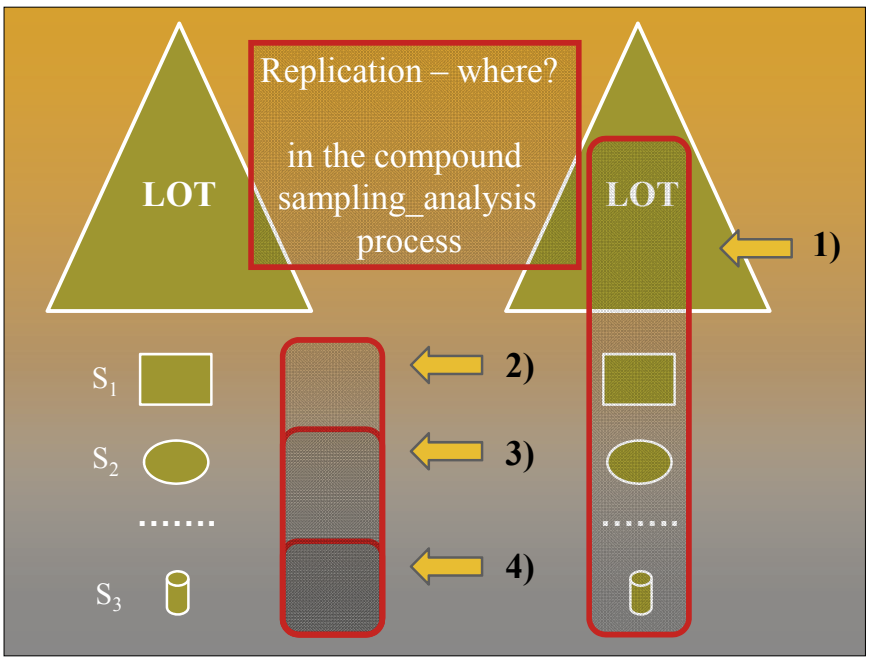

Figure 1. Replication can be performed at many stages in the full lot-to-aliquot pathway, but which is the most realistic situation pertaining to the general operations not **only** in the analytical laboratory? It turns out that all replication must meaningfully start "from the top".

equally reasonable to include some, or all, of the "sample preparation" procedures in the replication as well, because these part-operations cannot necessarily be performed in completely identical fashion. This effect should then also be repeated, say 10 times (stages 4 and/or 5 above) in order to acquire a measure of its variance contribution.

But having broadened the horizon this far, it is a logical step to follow up with still further realistic perturbations of the measurement process, which broadly means including also the tertiary, secondary and in the full measure of things, even also primary sampling errors in the replication concept. Why? Because these are *de facto* uncertainty contributions that will have been in play for any-and-all analytical aliquot, ever subjected to measurement! Following the full impact of the *Theory of Sampling (TOS)* and its detailed treatment of the phenomenon of *heterogeneity*, it is clear that the only complete "sampling-and-analysis" scenario that is guaranteed to include **all** uncertainty contributions must start with replication of the primary sampling ("replication from the top"). Any less comprehensive replication scenario is bound to be incomplete.

Repeating the primary sampling, again say 10 times (preferentially more when needed), means that each of 10 individually sampled primary samples is being subjected to an identical protocol that governs **all** the ensuing subsampling (mass-reduction), sample handling and preparation stages and procedures in the laboratory. From the logic of this full representativity pathway, "from lot-to-analytical aliquot", this is the only procedure incorporating the complete ensemble of uncertainties and errors encountered of whatever nature (sampling, handling, preparation, presentation). The point is that for each replicated primary sample, all potential errors will be manifested *differently* ten individual times giving rise to an accumulated variance which is the most realistic estimate of the **total** measurement uncertainty (MU).[2] In particular this estimate is bound to include the full sampling error effects (TSE), which will often dominate.

In clear contrast, starting at **any** other of the levels in the list above, stages 2–6 will guarantee an incomplete, inferior

SAMPLING COLUMN

TSE + TAE estimation, which is structurally destined to be too low, i.e. unrealistic.

Should one nevertheless feel compelled to "shortcut" the full replication procedure by not starting "from the top", one is **mandated** to describe the rationale behind this choice and to provide a **full report** of what was in fact done, lest the user of the analytical results has no way of knowing what was implicated in the umbrella term "replication". "Users" and decision makers, acting on the analytical data, do not like to be kept in the dark.

Undocumented or unexplained application of the term "replicate experiments" (or "repeated experiments") has been the source of a significant amount of unnecessary confusion in the past. Many times s^2(TAE) has simply been *misconstrued* to imply s^2(TSE + TAE), a grave error, for which *someone* or *somebody* (or some ill-considered, incomplete protocol) is responsible. But we are here not interested in pointing fingers at any entity (private or legal); it suffices to stop continuing such practice.

The above scenario illustrates an unfortunate responsibility compartmentalisation, which is sometimes found in scientific, industrial, publishing or regulatory contexts:

"The analyst is not supposed to deal with matters *outside* the laboratory (e.g. sampling)"

"This department is *only* charged with the task of reducing the primary sample to manageable proportions, as per codified laboratory's instructions"

"Sampling is automated and carried out by process analytical technology (PAT) sensors; there is no sampling issue involved here"

"I am not responsible for sampling, I only analyse/model the *data*"

... and similar *excuses* for not seeing the complete measurement uncertainty context. All too often the problem belongs to "somebody else", with the unavoidable result that the problem does not receive further attention. Therefore this stand ("not our responsibility") is always potentially in danger of being perpetuated and if so "replicate analysis" will still take its point of departure at stage 3 (maybe stage 2), but never from stage 1, the primary sampling stage. This is not an acceptable situation. There are many occasions in which authors, reviewers and even editors have missed cracking down with the necessary firmness on such demonstrable ambiguities regarding "replication", with the certain result that the reader is not able to understand what was intended, nor what was indeed carried out, because of incomplete descriptions in the "Method" sections of scientific publications and technical reports. The issue is therefore far from trivial, indeed grave errors are continuously being committed. But rather than address the obvious first question: who is responsible, the way forward shall here be constructive. The focus shall be on ways and means to put an effective end to the confusion surrounding the replication issue, and indeed put it to good use instead.

Quantifying total empirical variability—the replication experiment

Above was outlined how a realistic estimate of the *total* TSE + TAE, a replication experiment (RE) must always start "from the top". This is where replication starts, be this primary sampling in nature, in the field, sampling in the industrial plant, or it can be sampling of any target designated as the primary lot (examples follow below).

Figure 2 shows the scenario in which an avid sampler is facing a large lot with the objective of establishing a realistic estimate of the average lot concentration for one (or more) analytes. It is abun-

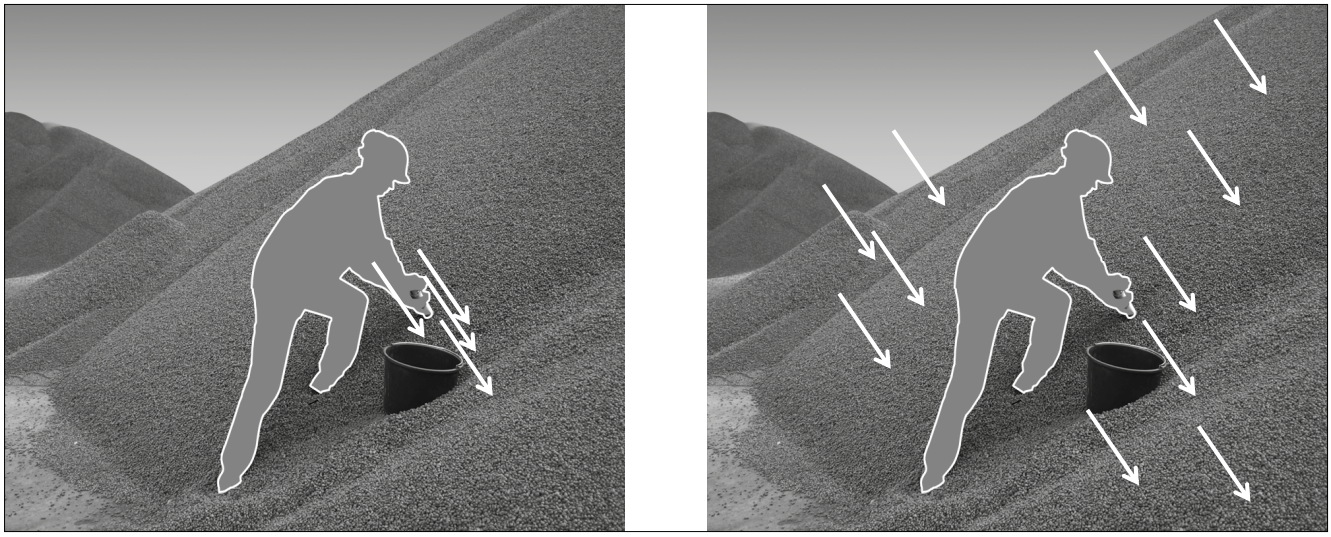

Figure 2. A primary sampler approaching a significantly heterogeneous lot with a grab sampling RE approach, but deployed with two very different coverage footprints. The left side realises the RE on an irrationally narrow footprint in relation to the full geometrical scale of the lot. The right side attempts to take account of the (hidden) lot heterogeneity by employing a wider footprint as a basis for the RE. These alternative scenarios will result in different relative sampling variability estimates because of the different lot heterogeneities covered. (N.B. neither of these primary sampling procedures succeeds to sample the interior of the lot, so both are not honouring the fundamental sampling principle (FSP).

dantly clear that a single grab sample stands no chance of ever being able to do this job because of the intrinsic distributional heterogeneity of the lot. It does not matter whether the lot is small, intermediate or large, the point being that this intrinsic heterogeneity is *unknown* at the moment of routine sampling. The sampler therefore has no other option than to act as if it is significantly large. There is no problem assuming this rational stance, the TOS furnishes all necessary governing principles and practical procedures and equipment assessment possibilities so as always to be able to deal with significant lot heterogeneity, e.g. Esbensen & Julius (2009).[3]

By deploying a RE, Figure 2 (right), the sampler now has access to a first estimate of the effective variability of the sampling procedure, but with TOS it is also clear that there is a grave breach of the fundamental sampling principle (FSP).

Relative sampling variability

It has been found useful to employ a general measure of the sampling variability as expressed by a RE, enter the *RSV*: the relative sampling variability.

The variability of any number of replications can be quantified by extracting and analysing the analytical results from a number of replicate primary samples. These specifically shall have the aim to *cover* the entire spatial geometry of the lot *as best possible*, i.e. spanning the geometrical volume of the primary lot in an optimal fashion (given the circumstances), and calculating the resulting empirical variability based of the resulting analytical results a_S. Often a relatively small number of primary samples may suffice for a first survey, though never less than 10. It is essential that the sampling operations are fully realistic replications of the standard routines, i.e. they shall **not** be extracted at the same general location, Figure 2 (left), which would only result in a *local* characterisation not at all able to relate to the effects of the full lot heterogeneity. What is meant here is that the successive primary sampling events shall take place at other, *equally likely* locations where the primary sampling is to be replicated. The RE shall be carried out by a fixed procedure that specifies precisely how the following sub-sampling, mass reduction and analysis are to be carried out. It is essential that both primary sampling as well as all sub-sampling and mass-reduction stages and sample preparation is replicated in a completely identical fashion in order not to introduce artificial variability in the assessment.

Note that when these stipulations are followed it is possible to conduct a RE for any sampling procedure, for example a grab sampling vs a composite sampling procedure.

It has been found convenient to employ a standard statistic to the results from a RE. The relative coefficient of variation, CV_{rel} is an informative measure of the relative magnitude of the standard deviation (*STD*) in relation to the average (X_{avr}) of the replicated analytical results, expressed as a %:

$$CV_{rel} = \left[\frac{STD}{X_{avr}}\right] \times 100 = RSV \qquad (1)$$

RSV is called the relative sampling variability (or relative sampling standard deviation). *RSV* encompasses all sampling and analytical errors combined as manifested by a minimum 10 times replication of the sampling process being assessed. *RSV* therefore measures the total empirical sampling variance influenced by the specific heterogeneity of the lot material, *as expressed by the current sampling procedure*. This is a crucial understanding. There can be no more relevant summary statistic of the effect of repeating the full lot-to-aliquot pathway procedures (10 or more times) than a RE-based *RSV*.

In the last decade there has been a major discussion in the international sampling community as to the usefulness of a singular, canonical *RSV* threshold; opinions have been diverse. In the last few years a consensus has emerged, however, that *indicates* a general acceptance threshold of 20%. *RSV*s higher than 20% signify a too-high sampling variability, with the consequence that the sampling procedure tested must be improved so as better to counteract the inherent heterogeneity effects in the lot material. Should one elect to accept a *RSV* higher than 20% this shall have to be justified and made public to ensure full transparency for all stakeholders.

The usefulness of the *RSV* measure cannot be underestimated. For *whatever*

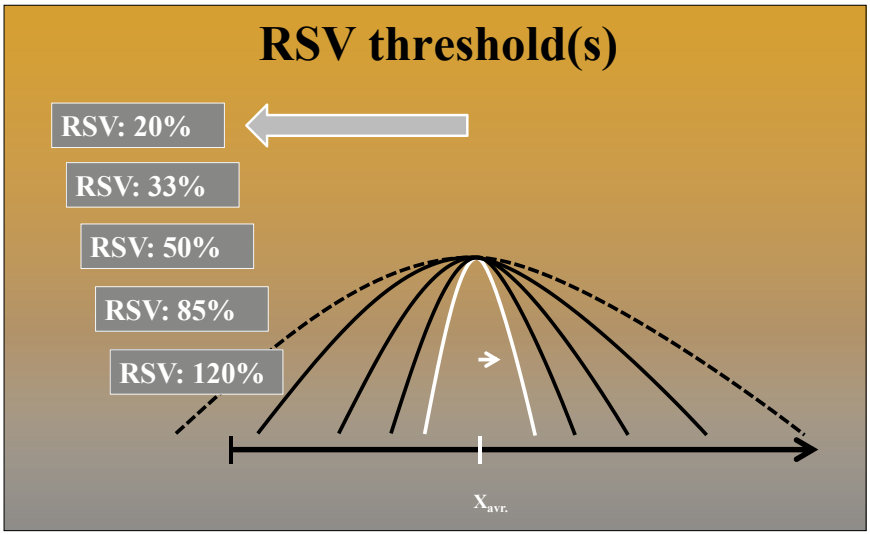

Figure 3. Schematic illustration of replication experiment thresholds *RSV*, e.g. 20%, 33%, 50%, 85% and 120%. Very large relative standard deviations (higher than approximately 85%), when interpreted as representing a standard normal distribution, apparently give rise to negative concentration values. This has no physical meaning, however, and need not cause any untoward worry; these are but model fitting artefacts, of no practical consequence. The essential information for the sampler is manifest already when *RSV* transgresses >20%, i.e. when the sampling procedure is operationally too variable and **must** be improved upon (TOS).

SAMPLING COLUMN

lot material, sampled by *whatever* procedure, the specific lot/procedure *combination* can be very quickly assessed. There are no untoward practicalities involved which might militate against performing a RE assessment; indeed anybody can perform RE assessment on any sampling procedure, or for any sampling equipment etc. It should never be possible to argue for, or against, a specific sampling procedure without a transparent quantitative assessment. RE numbers speak for themselves. The "difficult" issue of sampling is put on a fully understandable, and very simple operational basis—the RE.

Based on an extensive practical experience over 50 years from many applied sectors and fields within science, technology and industry, there are very many cases on record in which the 20% threshold is exceeded (not infrequently by significant deviations); but there are also an important number of cases in which the existing procedure is vindicated. A few illustrative examples are given below. But first: what information is residing in a simple *RSV* level?

Figure 3 illustrates how *STD* is expressed as a fraction of the general level quantified by X_{avr}. In this illustration the white distribution has a *STD* which is exactly 20% of X_{avr}. Also indicated are cases where the empirical *STD* forms, e.g. 33%, 50%, 85%... The issue clearly is, at what %-level is one no longer comfortable with the quantification resolution, e.g. for *RSV*=50% the signal-to-noise ratio is 1 : 1 only, likely not an acceptable situation under any accounts.

The canonical *RSV* threshold, 20%, serves as a general indication only in the case where *nothing* is known *a priori* as to the heterogeneity of the material involved. Materials and material classes certainly exist that may merit a higher, or a lower, threshold, for which the proposed general *RSV* value can, of course, no longer serve. For such cases, a material-dependent quantification can be developed, dependent upon the sampler's own competence and diligence. The mandate in the sampling standard DS 3077[4] is clear: **all** analytical results shall be accompanied by an appropriate *RSV*, voluntarily described and reported in full.

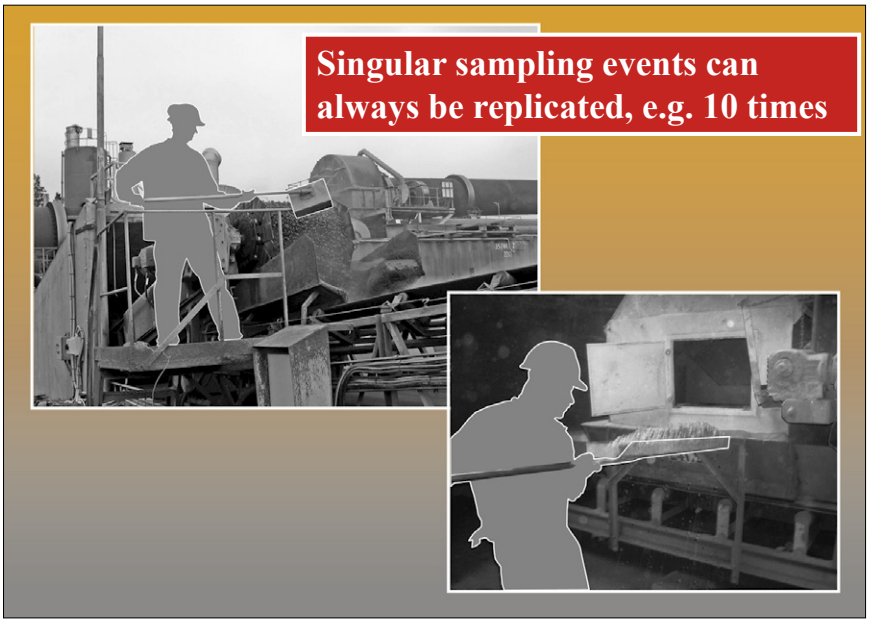

Figure 4. Examples of replication experiments (RE) that are easily set up. On the left is a dynamic process sampling situation, at the right sampling from a stationary lot. Both sampling scenarios can be assigned an objective *RSV* quality index. In order that no misunderstanding may occur, it is only necessary to perform a proper, calibrating RE **once**, as part of surveying and characterising the intrinsic heterogeneity of a specific lot material.

While it is acceptable to level criticism against the suggested threshold (20%), this also entails the obligation to perform empirical due diligence in the form of a RE. Recent industrial, scientific and technological history is flush with examples of major surprises brought about by such simple replication experiments and their attendant *RSV*. It is either the intrinsic material heterogeneity which is underestimated or, at other times, the sampling procedure turned out to be much less universal than assumed.

The purpose of a RE is often to assess the validity of an already existing sampling procedure. In practice, the RE can only perform and test a current sampling procedure as it *interacts* with a specific lot material. *Should* a *RSV* for this exploratory survey exceed the canonical, or case-specific, threshold, the need for complete fulfilment of the TOS has been documented and is therefore mandated, no exceptions allowed. There may be good reasons to start validation by testing an existing sampling procedure—there is always the *possibility* it may turn out to fall below the pertinent threshold, and thus be acceptable as is. But in all other cases, TOS-modifications must be implemented, no exceptions.

One can therefore view *RSV* as a flexible and relevant sampling procedure *quality index*, scaled with the inherent heterogeneity encountered. *RSV* is particularly useful for initial characterisation of sampling from *stationary lots*, while it is much more customary to use a dynamic, process sampling augmented approach, called *variographics* when sampling from dynamic lots. *RSV* and *variographics* are closely related approaches fundamentally quantifying the same heterogeneity; the latter approach is much more powerful, however, due to the fact of its more elaborate experimental design which allows full decomposition of GEE, see, for example, References 5–8. Variographic heterogeneity characterisation of dynamic lots is the subject of a later column in this series.

All examples described above pertain to issues related to sampling and other error contributions **before** analysis. It is noteworthy that some analytical procedures can have significantly large TAE, e.g. of the order of 10–20% or more,

which is then already factored into the empirical *RSV* level. The principle issues from the few examples given here can be generalised to many other material and lot types. The GEE=TSE+TAE issues are identical for all lot systems.

The following examples illustrate how a specific sampling equipment can be assessed with respect to several different materials (with specific heterogeneities), which may result in both pass and fail.

RE is a general facility that can in fact be deployed at all stages in the lot-to-aliquot pathway, i.e. also a stages later than the primary sampling stage. If the objective were to assess and compare the two splitters in Figure 6 specifically, the RE may well be initiated at this sub-sampling stage directly (in such a case it is of course still critical to add the sampling error effects from the preceding stages in the final evaluation).

The replication experiment (RE) is a powerful and highly versatile sampling/analysis quality assessment facility that can be deployed with great flexibility. It is necessary to be fully specific as to what is meant by "replication" in the situation at hand, i.e. at what stage in the lot-to-analysis pathway is replication to commence. We shall have occasion to employ replication experiments many times in these columns.

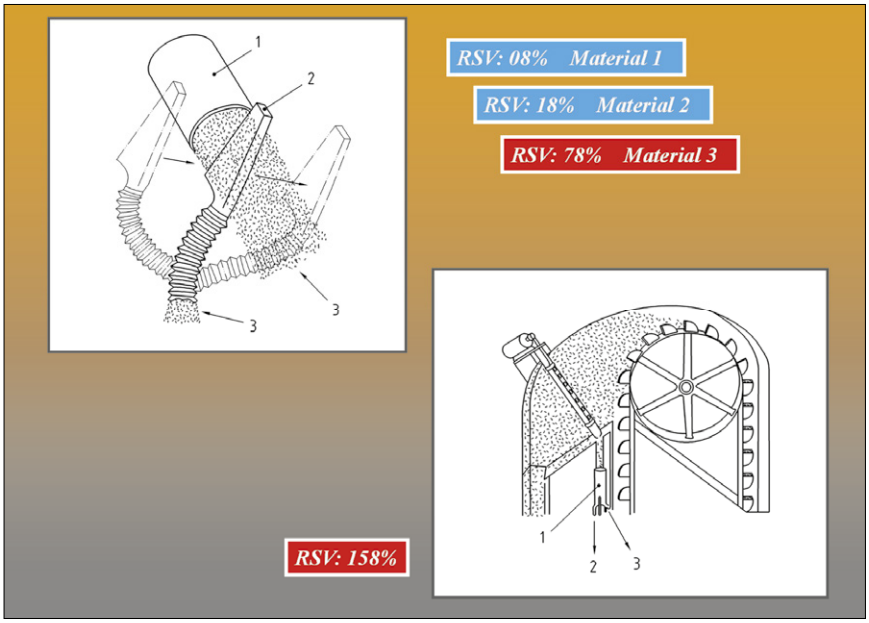

Figure 5. Upper left: primary process sampler assessed for three different materials, one of which does not pass the test of the dedicated RE (*RSV*=78%). Lower right: a complex primary sampler being subjected to a RE with the distinctly worrisome result of *RSV*=158%. N.B. illustrative examples only, no specific sampler is endorsed, nor renounced. Samplers are sketched only in order to illustrate how RE may be used for quantitative assessment.

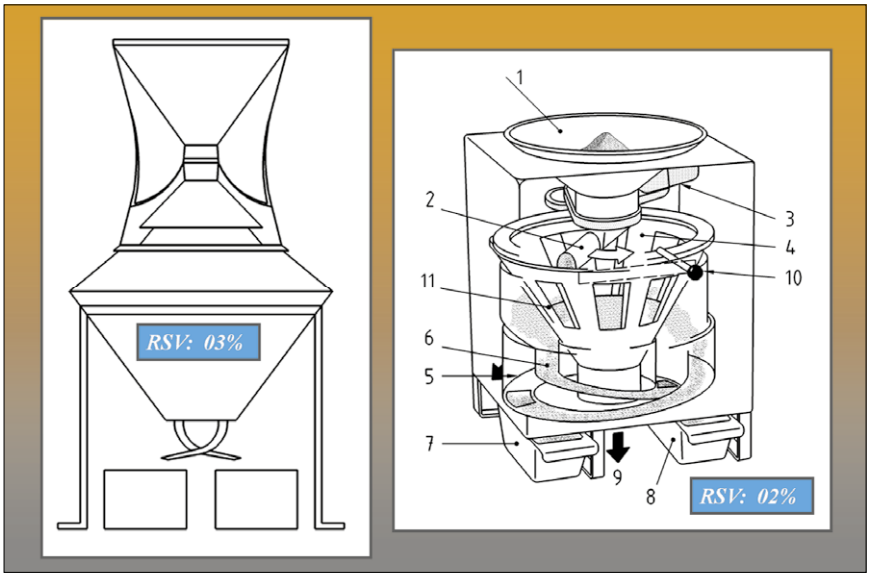

Figure 6. Two laboratory equipment (splitters) subjected to RE assessment, showing highly satisfactory quantitative results. N.B. illustrative examples only, no specific sampler is endorsed, nor renounced. Samplers are sketched only in order to illustrate how RE may be used for quantitative assessment.

Notes and references

1. Taguchi approach: http://en.wikipedia.org/wiki/Taguchi_methods
2. Issues related to the concept of Measurement Uncertainty (MU), which too often in practice only covers the parts of the analysis process that can be brought under direct laboratory control (while in its full definition purports to cover the entire sampling-handling-analysis pathway, are treated in: K.H. Esbensen and C. Wagner, "Theory of Sampling (TOS) versus Measurement Uncertainty (MU) – a call for integration", *Trends Anal. Chem.* **57,** 93–106 (2014). doi: http://dx.doi.org/10.1016/j.trac.2014.02.007
3. K.H. Esbensen and L.P. Julius, "Representative sampling, data quality, validation – a necessary trinity in chemometrics", in *Comprehensive Chemometrics*, Ed by S. Brown, R. Tauler and R. Walczak. Elsevier, Oxford, Vol. 4, pp. 1–20 (2009).
4. K.H. Esbensen (chairman taskforce F-205 2010-2013), *DS 3077. Representative Sampling—Horizontal Standard*. Danish Standards (2013). http://www.ds.dk
5. K.H. Esbensen, A.D. Román-Ospino, A. Sanchez and R.J. Romañach, "Adequacy and verifiability of pharmaceutical mixtures and dose units by variographic analysis (Theory of Sampling) – A call for a regulatory paradigm shift", *Int. J. Pharm.* **499,** 156–174 (2016). doi: http://dx.doi.org/10.1016/j.ijpharm.2015.12.038
6. K.H. Esbensen, C. Paoletti and P. Minkkinen, "Representative sampling of large kernel lots – I. Theory of Sampling and variographic analysis", *Trends Anal. Chem.* **32,** 154–164 (2012). doi: http://dx.doi.org/10.1016/j.trac.2011.09.008
7. K.H. Esbensen, C. Paoletti and P. Minkkinen, "Representative sampling of large kernel lots – III. General Considerations on sampling heterogeneous foods", *Trends Anal. Chem.* **32,** 178–184 (2012). doi: http//dx.doi.org/10.1016/j.trac.2011.12.002
8. P. Minkkinen, K.H. Esbensen and C. Paoletti, "Representative sampling of large kernel lots – II. Application to soybean sampling for GMO control", *Trends Anal. Chem.* **32,** 165–177 (2012). doi: http://dx.doi.org/10.1016/j.trac.2011.12.001