The Replication Myth 2: Quantifying empirical sampling plus analysis variability

Column Editors: Kim H. Esbensen,[a] Paul Geladi[b] and Anders Larsen[c]

[a]**Geological Survey of Denmark and Greenland (GEUS), Denmark**
[b]**Swedish University of Agricultural Sciences, Umeå, Sweden**
[c]**Q-Interline, Denmark**

I n the previous column, it was stated that in order to furnish a realistic estimate of the complete Total Sampling Error (*TSE*) + Total Analytical Error (*TAE*), a Replication Experiment (RE) must always start "from the top", i.e. it is the primary sampling which is to be replicated. The primary sampling may be field sampling, sampling at the industrial plant or it can be sampling of any lot designated as the primary lot (examples follow below). In this fashion, and *only* in this fashion, will all sampling errors be represented (primary sampling, sub-sampling, mass reduction errors associated with sample preparation etc.) in ten (or more) individual manifestations—to which is added the total analytical error, *TAE*.

Figure 1 illustrates 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 of one (or more) analytes. It is abundantly clear that a single grab sample stands virtually no chance of ever being able to do this job because of the intrinsic heterogeneity of the lot. It does not matter whether this is small, intermediate or large; the point is 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 and the Theory of Sampling (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.[1] Trying out a single grab sample will result in one analytical result only, with no possibility whatsoever to assess the relevance (and representativity) thereof. By deploying a proper RE [Figure 1(b)], the sampler now has access to an estimate of the effective variability of the sampling procedure. From the geometry it is clear that the lot heterogeneity will be better quantified in the case pictured in Figure 1(b). The ultimate issue of representativity cannot by itself be determined without recourse to TOS—from which it transpires that composite sampling must be employed. The issue of grab sampling vs composite sampling will be dealt with in a later Mythbuster column. Here we present a quantitative way to express and to compare the results of replication experiments.

Relative sampling variability (*RSV*)

It has been found useful to employ a general measure of the sampling variability as expressed by a RE, i.e. the *RSV*—Relative Sampling Variability.

The variability of any number of replications can be quantified by extracting and analysing a number of replicate primary samples. These specifically shall aim to *cover* the entire spatial geometry of the lot in the best way 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 on 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 (N.B. preferably more). It is essential that the primary sampling operations are fully realistic replications of the standard routines, i.e. they shall *not* be extracted at the same general location [Figure 1(a)] since this would result in a local characterisation which is surely too conservative with respect to the full lot heterogeneity, which is the heterogeneity encountered by any new sampling. Instead, the chosen number of replications shall *cover* the full geometrical domain of the lot [to the best possible extent considering the number of replications available, Figure 1(b)]. What is meant here is that the successive primary sampling events shall take place at other, *equally likely* locations if the primary sampling was to be replicated
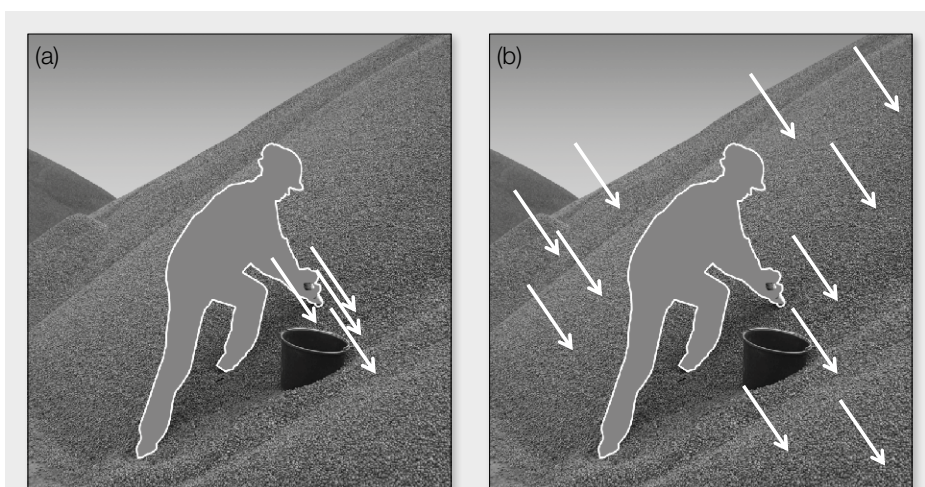

Figure 1. A primary sampler approaching a significantly heterogeneous lot with a grab sampling RE approach but deployed with two very different coverage footprints. (a) Realises the RE on an impossibly narrow footprint in relation to the full geometrical scale of the lot. (b) Attempts to take account of the (hidden) lot heterogeneity by employing a wide footprint as a basis for the RE. These alternative scenarios will result in different *RSV* estimates because of the different lot heterogeneities covered. (N.B. neither of these primary sampling procedures samples the *interior* of the lot, so they both therefore pose severe issues with respect to representativity. How to resolve this issue is treated in the TOS literature).

from the start. The replication experiment shall be carried out by a fixed procedure that specifies precisely how the subsequent sub-sampling, mass reduction and analysis are to be carried out. It is essential that primary sampling as well as all sub-sampling and mass-reduction and sample preparation stages are replicated in a completely identical fashion in order not to introduce artificial variability in the assessment.

It has been found convenient to employ a standard statistic to the results from a replication experiment. 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 percentage:

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

RSV (%) is called the Relative Sampling Variability (or relative sampling standard deviation).

RSV (%) encompasses all sampling and analytical errors combined on the basis of a minimum 10 times replication of the sampling process being assessed. RSV (%) measures the total empirical sampling variance influenced by the specific heterogeneity of the lot material, as expressed by the current sampling procedure. RSV is comprised of both the primary, secondary and tertiary sampling errors, including all errors incurred by mass reduction—as well as the TAE. There is 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 canonical RSV threshold; opinions have been diverse along the way. In the last few years, a consensus has, however, emerged that suggests a general acceptance threshold of 20%. RSV (%)'s higher than 20% signify a sampling variability which is too high, indicating that the sampling procedure tested must be improved. Should one elect to accept a RSV higher than 20%, this should be justified and made public to ensure full transparency for all stakeholders.

The usefulness of a general RSV threshold of 20% cannot be underestimated. For *whatever* lot material, sampled by *whatever* procedure, the specific lot/procedure combination can be quickly assessed by a simple RE. 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 is simplicity itself to carry out a RE and from this moment it will never be possible to try to argue for, or against, a specific 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 considerable practical experience over 50 years or so 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 (significantly in some cases) and, of course, 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 2 illustrates how STD is expressed as a fraction of the general level quantified (X_{avr}), i.e. the white distribution has a STD which is exactly 20% of X_{avr}. Also indicated are cases in which the empirical STD forms, e.g. 33%, 50%, 85% …, with respect to X_{avr}. 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 and this is not an acceptable situation under any circumstances.

The canonical RSV threshold of 20% serves as a general indication in cases about which nothing is known *a priori* of the heterogeneity of material. Materials and material classes that may merit a higher or lower threshold certainly exist, and in these cases the proposed general RSV (%) value
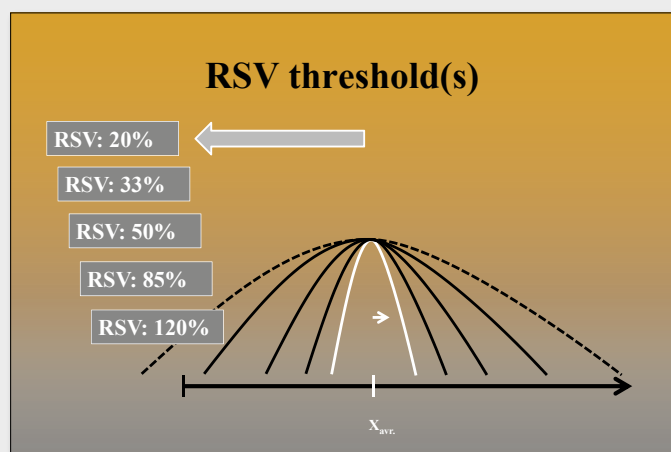

Figure 2. Schematic illustration of replication experiment thresholds RSV, e.g. 20%, 33%, 50%, 85% and 120%. Very large relative standard deviations (higher than approximately 85%), when fitted to a standard normal distribution, apparently give rise to *negative* concentration values. This, however, has no physical meaning nor should it 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 exceeds 20%, i.e. when the sampling procedure is operationally too variable and *must* be improved upon (TOS).

Figure 3. Examples of replication experiments (RE) easily set up. At left is shown a dynamic process sampling situation, at 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.

shall not necessarily be applied. For such cases, a material-dependent quantification, *RSV* (%), can be developed, dependent upon the sampler's own competence and diligence. The only stipulation is that all deviations from the general threshold level shall be individually justified, described and reported in full. Thus it is, for example, acceptable to quibble about the suggested threshold (20%), just as long as one performs due diligence in the form of an alternative RE. Recent industrial, scientific and technological history is full of examples of major surprises brought about by such extremely simple replication experiments. It is often typically the intrinsic material heterogeneity which is underestimated—at other times the sampling procedure turned out to be much less universal than expected.

Quality control of a replication experiment is strongly influenced by the degree to which *realistic* and relevant compositional and spatial coverage of the entire lot has been achieved. It is fully possible to try to dodge this issue by deliberately replicating the primary sampling based on a very restricted "local footprint" only [Figure 1(a)]. However, it is clear that such intentions and behaviour fall outside the characteristics pertaining to fully responsible operators and scientists.

The purpose of a RE is often to assess the validity of an already existing procedure. In practice, the RE is meant to test a current sampling procedure as it *interacts* with a specific lot material. A proper RE will include all error effects from whatever heterogeneity, sampling and analysis errors involved. Should the *RSV* for this exploratory survey exceed the canonical, or case-specific, threshold, the need for complete fulfilment of the TOS has been documented and 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 when this is not the case, TOS-modifications must be implemented. One should therefore view *RSV* as a flexible and relevant sampling procedure quality index, scaled by 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 issues, i.e. *TSE*; the latter is much more powerful, however, due to its more elaborate experimental design which allows full decomposition of Global Estimation Error (*GEE*).[2–4]

All the present illustrative examples pertain to issues related to sampling error contributions before analysis. It is noteworthy that *some* analytical procedures can have significantly large *TAE*, e.g. of the order of 10% or more, which is then already factored into the empirical *RSV* level. The principle issues from the few examples given here can easily be generalised to very many other material and lot types. The *GEE = TSE + TAE* issues are identical for all systems. Therefore, a RE can be useful in all situations in which there is no *a priori* knowledge of the individual magnitude of *TSE* (*TAE* is quite often known).

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

As illustrated, the general RE is a highly versatile facility that can be deployed at all stages in the lot-to-aliquot pathway, i.e. also at stages later than the primary sampling stage. If the objective was to assess and compare the two splitters in Figure 5 specifically, the RE may well be initiated at this sub-sampling stage directly (in such a case it is of course critical to add the sampling error effects from the preceding stages in the final evaluation as well).

Summing up Replication Myth columns 1 and 2: always 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
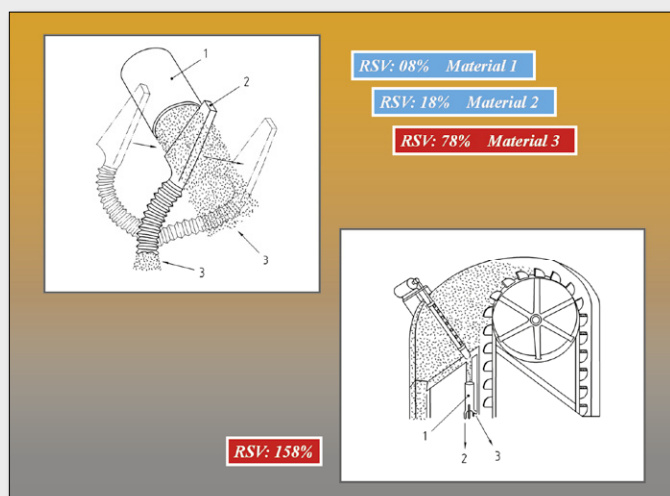
Figure 4. Upper left: primary process sampler assessed for three different materials, one of which does not pass the test of the dedicated RE (*RSV* = 78%). Lower right: a complex primary sampler suggestion being subjected to a RE with the worrying result of *RSV* = 158%. N.B. illustrative examples only, no specific sampler is endorsed, nor renounced. Samplers are sketched only in order to illustrate how RE *may* be used for quantitative assessment.
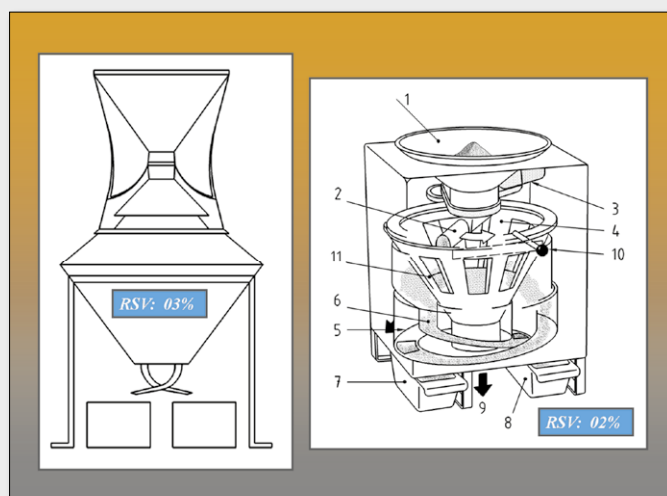
Figure 5. Two laboratory samplers (splitters) subjected to RE assessment, showing highly satisfactory quantitative results. N.B. illustrative examples only, no specific sampler is endorsed, nor renounced. Samplers are sketched only in order to illustrate how RE *may* be used for quantitative assessment.

pathway is replication to commence. Make liberal use of the Replication Experiment facility (RE), specifically establishing a quantitative basis for assessment in the form of *RSV* (%).

The RE is a powerful sampling/analysis quality assessment facility that can be deployed with great flexibility.

References

1. K.H. Esbensen and L.P. Julius, "Representative sampling, data quality, validation—a necessary trinity in chemometrics", in *Comprehensive Chemometrics*, Vol. 4. Ed by S. Brown, R. Tauler and R. Walczak. Elsevier, Oxford, pp. 1–20 (2009).

2. 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,** 154–165 (2012). doi: 10.1016/j.trac.2011.09.008

3. 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,** 166–178 (2012). doi: 10.1016/j.trac.2011.12.001

4. 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,** 179–184 (2012). doi: 10.1016/j.trac.2011.12.002