9. Replication—replicates—but of what?

This chapter provides an overview of an issue which has not always received its fair attention, the question of "replication", exemplified by the following examples that can be found frequently in the literature: i) replication of sampling; ii) replication of samples); iii) replication of measurements (analysis). What is meant by *replication* here? Are these "replications" identical? The commonly implied connotation is that of a beneficial averaging carried out with the help of replicates. There are, however, many uncertainties and imprecise assumptions involved when considering averaging, averaging of what exactly? This issue needs careful consideration before data analysis can be performed appropriately.

9.1 Introduction

From the discipline of experimental design (design of experiments, DOE, chapter 11), comes a well-organised strict understanding and terminology for "replicate measurement", because of the rigidly controlled situation surrounding the actual design. For example, in the situation of chemical synthesis influenced by several experimental factors (say, temperature, pressure, concentration of co-factors), it is easy to understand what a replicate measurement means: repeat the synthesis experiment under **identical** conditions for these controllable factors and replicate the outcome measurement, for example, the yield. By definition of a designed experiment, care has been taken to randomise all other potential factors, in which case the variance of the experimental results, be it small or large, is supposed to furnish a measure of the "Total Analytical Uncertainty". Upon reflection, however, this variance also tells of the uncertainty contributions stemming from other influences, for example, from small-scale sampling of reactants involved, which may not necessarily represent "homogeneous stocks", but more likely are of uniform composition only, see chapter 3. Added uncertainty contributions may also arise from resetting the experimental setup, i.e. to what precision can one "reset" parameters such as temperature, pressure or concentration levels of co-factor chemical species after having turned the setup off and cleaned all the experimental equipment (perhaps even waiting until next week) before "replicating"? However, such uncertainty contributions are usually considered insignificant because of the "controllable" situation attending DOE.

There are, however, many other scenarios preceding data analysis that far from parallel this nicely bracketed situation of a controlled experiment. Indeed, most data sets do not originate from within the complacent four walls of an analytical laboratory only, but from sampling of heterogeneous systems and processes from all of science, technology and industry. What are described below constitute the opposing end of a full spectrum of possibilities in which the researcher/data analysts must recognise as *significant* sampling, handling and preparation errors in addition to Total Analytical Error (TAE). The issue can most conveniently be organised around one key concept: What is meant by "replicate samples"? Note that one physical sample may end up being represented in a data matrix as several objects etc. with a very real danger of potential confusion.* To add to the complexities, replication may also be carried out at other levels, for example, related to sampling of multiple batches, lots or production units, sampling from different seasons or from different instruments, perhaps carried out by different operators. All these higher-order replication issues are discussed in more detail in chapter 8. For each of these scenarios, it is imperative that the reader is offered full disclosure of the stage, level and intensity of replication.

* This chapter is intended to deal comprehensively with the confusion surrounding all issues of "replication". A general argument will be presented covering the most common and also less common scenarios.

Upon reflection, this issue will appear more complex than what may seem the case at first sight, indeed it merits careful definition, contemplation and a strict terminology among other reasons because it is also intimately related to the validation issue presented in chapter 8. There has been much confusion (due to unawareness and sometimes neglect) because of far too vague or incomplete definitions of *what* it is exactly that is replicated.

As the case in point, what is meant by "replicated samples"?

With reference to TOS, it will be appreciated that "replication" can concern (at least) the following alternative scenarios:

Stage 1: Replication of the primary sampling process (all the way to analysis), with due regard to the possible effects of time, sequence, raw material variation etc.

Stage 2: Replication starting at the secondary sampling stage (i.e. first mass reduction).

Stage 3: Replication starting with the tertiary sampling process (further mass reduction).

Stage 4: Replication starting with aliquot extraction and preparation (e.g. involving compaction or other problem-dependent operations, which all add variance when replicated).

Stage 5: Replication starting with aliquot instrument presentation (e.g. surface conditioning).

Stage 6: Replication of the analysis only (TAE).

The last situation logically corresponds to the term "replicate analysis". But does this mean that the aliquot (the vial) stays in the analytical instrument all the time while the analyst "presses the button" repeatedly, say 10 times? Possibly—this would indeed correspond to TAE *sensu stricto,* but it may seem equally relevant to extract the vial and insert it in the instrument repeatedly, allowing normal temperature variations to influence TAE because this is a more *realistic* repetition of the between-samples situation than repeating measurements on one static sample housed in the analytical instrument without replacement. This is a first foray into "Taguchi thinking". † But to another analyst, it may perhaps appear equally reasonable also to include some, or all, of the "sample conditioning/preparation" variations in the replication scheme, for the same reason: to be more realistic. For which reason, such perturbations should then logically also be repeated 10 times (stages 4 and/or 5 above).

Having opened up this avenue, it now seems an unavoidable logical step to follow up with further, equally relevant and realistic perturbations of the circumstances surrounding "analysis", and in fact include also the tertiary, secondary and ultimately the primary sampling errors in the replication concept. Following the full impact of chapters 3 and 8, it is clear that the only **complete** "sampling-and-analysis" scenario, which is guaranteed to include **all** possible uncertainty contributions to the total Global Estimation Error (GEE), is the one that starts with replication of the primary sampling method ("replication, from the top").

Repeating the primary sampling, say, 10 times, each sample being subjected to the exact same protocol governing all the ensuing sub-sampling (mass-reduction), sample handling and preparation, is the only procedure that allows the full set of uncertainties and errors‡ to be reliably manifested more than once,

† Taguchi approach: http://en.wikipedia.org/wiki/Taguchi_methods

‡ "There is always an *uncertainty*, regardless of how small it is, between the true, unknown content a_L of the lot *L* and the true, unknown content of the sample, a_S. … tradition has established the word 'error' as common practice, though it implies a mistake that could have been prevented, while statisticians prefer the word 'uncertainty' which implies no responsibility. However, in practice, as demonstrated in the Theory of Sampling, there are both sampling errors, and sampling uncertainties. Sampling errors can easily be minimised, while sampling uncertainty for a pre-selected sampling protocol in inevitable. …. Because the word 'uncertainty' is not strong enough, the word 'error' has been selected as current usage in the Theory of Sampling, making it very clear it does not necessarily imply a sense of culpability"; quoted from Pitard [1] p. 33 who graciously informs that this statement rightly originates with Pierre Gy (in a monograph written in French, 1967).

i.e. this approach is the only **fully** realistic replication of **all** the elements in the sampling-and-analysis pathway compared to the routine workflow of typical laboratories. By contrast, starting at any other of the levels in stages 2–6 is guaranteed to result in inferior TSE + TAE estimations, which structurally will be guaranteed to be too low.

There is always an obligation for the analyst to describe the rationale behind the specific choice of a replication scheme and to fully disclose voluntarily exactly what was in fact replicated, else the user of the analytical data will be in the dark. Undocumented, unexplained (and sometimes even ill-understood) application of the term "replicates" (w.r.t. sampling, samples, sub-samples, aliquots?) has been the source of a significant amount of unnecessary confusion in analytical chemistry, statistics and chemometrics. However, many times the problem boils down to that s^2(TAE) simply has been *misconstrued* to imply the much larger s^2 (TSE + TAE), a grave error, for which *someone* must be responsible—but who? Who or what is the culprit in this context? More importantly, how can this be rectified?

The above scenarios illustrate the unfortunate compartmentalisation of responsibility, which rules the day in a wide swath of current laboratory, scientific and industrial contexts. Comments commonly heard are: "..the analyst is not supposed to deal with sampling *outside* the laboratory"; "...this department is *only* charged with the important task of reducing the primary sample to manageable proportions, as per the laboratory's instructions"; "...sampling is *automated*, there is no sampling problem here"; "...I am not responsible for sampling, I analyse *the data*!" and a legion of similar excuses for not seeing, or wanting to deal with, the complete "measurement uncertainty" issue. All too often the problem belongs to "somebody else".

If "excuses" like these continue to be allowed in practical experimental work, in technical guidelines and reports and in the scientific literature, there is a grave danger that this unfortunate stand will only be perpetuated: "Replication" will then mostly still take its point of departure at stage 3 (maybe stage 2, but almost never from stage 1), the primary sampling stage. Add to this a distinct lack of stringency on behalf of authors, reviewers and editors to focus—or be knowledgeable enough to be able to focus—and crack down on this enormous ambiguity regarding "replication". The issue is manifestly critical. Grave errors are still being committed, for which a universal remedy has not yet seen the light. This chapter intends put an end to this unfortunate issue. How? Well, the issue is no longer what is wrong, the issue rather is: *what can be done about it*? Indeed, *who* is going to do something about it? Well, it turns out that the answer is very close—it's **YOU**, the insightful and competent data analyst, by insisting on appropriate data quality precautions as well as relevant strategies for collection of data to ensure representative variation in the data to be analysed, which is a key point in all multivariate considerations.

9.2 Understanding uncertainty

The basic assumption underlying the application of multivariate analysis is that the measured data carry *relevant information* about the studied properties and experimental objectives. It is obvious that it matters very much whether the individual data in any matrix are fraught with measurement uncertainties proportional to s^2(TSE + TAE) or to s^2(TAE) alone, else data analytical interpretations and conclusions run the risk of addressing patterns, results and issues which are in reality "below the effective uncertainty level". Data quality needs to be quantifiable and how to achieve this is the purpose of this section.

With reference to chapter 3, most aspects related to the replication issue can be assigned to incomplete or too vague notions as to the role and influence of spatial heterogeneity, DH_L of the analyte in question. But with proper insight one will never neglect the most influencing of TSE contributions from primary, secondary and tertiary sampling, effects which must be determined by an empirical inquiry. Depending on the application, various other aspects of heterogeneity in the system under observation might also be instrumental, for example, temporal irregularity. What is confusing, and frustrating, is that in many cases the data analyst is, quite literally, miles away from the location where the issue originates. It is not, however, impossible, difficult, nor expensive to do something about this: the Replication Experiment can easily be carried out by the same personnel responsible for the primary sampling.

9.3 The Replication Experiment (RE)

The scene is now set for a remarkably powerful tool with which to deal effectively with all issues regarding TSE vs TAE. The Replication Experiment (RE) will be able to resolve all the issues that were raised above, and in chapters 7 and 8, regarding the operative influence(s) on the total measurement uncertainty, and at all relevant sampling stages. As luck would have it, the principle of a Replication Experiment is simplicity itself.

The quantitative effect of lot distributional heterogeneity (DH_L) *interacting* with a specific sampling process (i.e. any sampling process based on a pre-selected sample mass in a specific sampling plan using either grab sampling or composite sampling, or whatever) can be quantified by extracting and analysing a small number of *replicate primary samples* with the objective to "cover the spatial geometry, or the salient time span of the lot/process" as best possible, and from this to calculate the empirical variance of the resulting analytical results a_S. This procedure is termed a *Replication Experiment*.

A relatively small number of primary samples may often suffice, though never less than 10 if possible. The issue at hand is not only statistical, but more associated with the presumed lot heterogeneity. If DH_L is, or is suspected of being, significantly influential, it is senseless to be frugal w.r.t. the number of replicated primary samples for obvious reasons. It is suggested to use an informative replication index, **r**, to the replication experiment term, RE(**r**). Thus RE(**7**) can be meaningfully compared to RE(**12**) for example. The issue is not so much the exact **r** value, it is rather that the experimenter honours an obligation to report on what basis RE was physically carried out—e.g. in Figure 9.1 a manual stream sampling subjected to a RE(**10**). Note that this particular sampling operation may, for example, not necessarily be representative—in which case this is precisely the kind of insight that will be provided by a replication experiment.

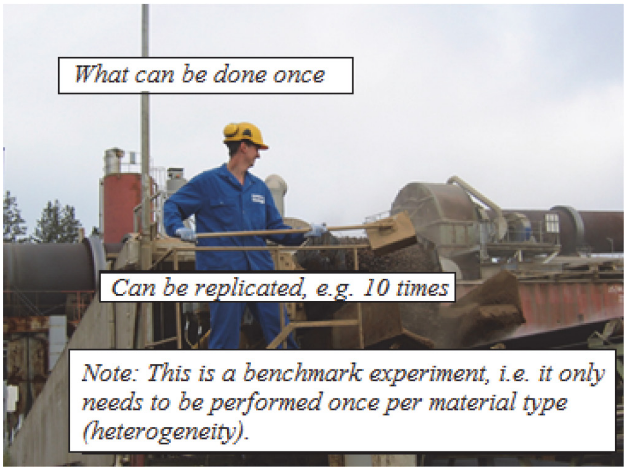
What can be done once
Can be replicated, e.g. 10 times
Note: This is a benchmark experiment, i.e. it only needs to be performed once per material type (heterogeneity).

Figure 9.1: A generic example of a specific primary sampling operation (in this case a manual sampling of a process stream) that can easily be replicated (for example 10 times), which is all that is needed for a Replication Experiment, RE(**10**).

The replication experiment must be governed by a protocol that specifies precisely how all procedural elements are to be carried out. It is essential that both primary sampling as well as all sub-sampling and mass-reduction stages, sample preparation etc. is replicated in a completely *identical* fashion. Obviously, it is preferable for a RE(**r**) that all Incorrect Sampling Errors (ISEs) have been eliminated, i.e. that only *correct sampling* is employed (TOS' preventive paradigm, refer to chapter 3). However, it is also feasible first to gauge an existing sampling-and-analysis procedure in which this requirement has not necessarily been fulfilled. The replication experiment will then *include* the pertinent error effects from these factors, i.e. include the adverse sampling bias effects. This will soon be seen as a major bonus, however, revealing the real practical power of RE(**r**).

It has been found convenient to employ a standard statistic to characterise results from the replication experiment. The relative coefficient of variation, CV_{rel} is an informative measure of the magnitude of the standard deviation (σ) in relation to the average ($\mathbf{X}_{avr}$) from a series of properly replicated analytical results, expressed in % (equation 9.1):

$$CV_{rel} = \left[\sigma / \mathbf{X}_{avr}\right] \times 100 = RSV \qquad (9.1)$$

When *RSV* is calculated from data originating from a replication experiment starting with the primary sampling, it is clear that it encompasses **all** sampling and analytical errors as manifested **r** times through the full "lot-to-analysis" pathway. *RSV* measures the total empirical sampling variance influenced by the specific heterogeneity of the lot material as expressed by the current sampling procedure. The core issue is that a properly designed *RSV* is not only a reliable (TSE + TAE) estimator, it simultaneously furnishes a quantitative measure of the effective heterogeneity of the lot, precisely *as manifested by the sampling procedure in use*. The specific *RSV* magnitude can be seen to be directly proportional to the effective total heterogeneity of a(ny) lot/material, because some form of sampling of the lot must be carried out in order to end up with analytical results. What could be more relevant than a proper quantitative characterisation of any practical "measurement situation" including the critical sampling component?

Since all sampling errors, at all scales from lot to analysis, are included for each replicated primary sample, it is certain that both the inconstant sampling bias (if present) as well as all Grouping and Segregation Error (GSE)-induced variability effects (always present to some degree, as are the Fundamental Sampling Error effects, FSE) are allowed to manifest themselves **r** replicated times. It follows that *RSV* provides a highly relevant expression of the effective **total** "measurement uncertainty". This has the desirable consequence that all sampling procedures may easily be put to the test, **if** a specific threshold with which to compare is at hand.

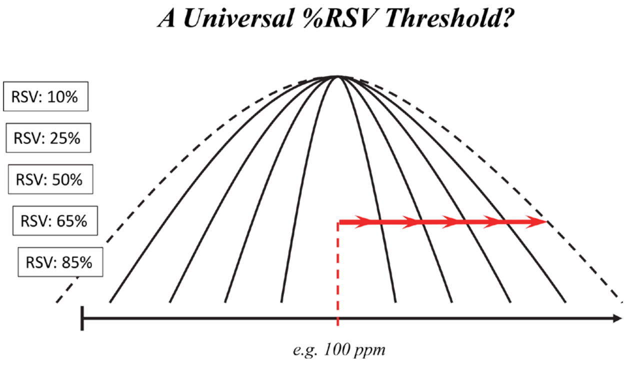

Figure 9.2: Examples of different empirical *RSV* magnitudes, expressed with respect to the pertinent average of the **r** analyses (for example 100 ppm as illustrated). The larger the *RSV* magnitude, the larger the spread of the **r** final analytical results realised through the RE(**r**). Is there a universal threshold?—see text.

From the TOS community, a general acceptance threshold has (reluctantly) been suggested as 20%, but as this is based on theoretical model understandings of FSE alone, an additional measure must be added allowing for the effects also from GSE for all types of significantly heterogeneous materials. Thus, a *RSV* which is higher than, say, 30% signifies an unacceptably high sampling variability—with the mandate that the sampling procedure must be *improved*. Remember, however, that such a general threshold issue is in reality passing judgement over what is strongly problem-dependent (materials heterogeneity may well be so that a higher threshold is warranted, but there also exist many types of materials for which a lower threshold than 20% is entirely appropriate§). Be all this as it may, for the data analyst, it is sufficient to be conversant enough with these matters to demand that some form of *RSV threshold* must be available, else data analysis operates with a much too uncomfortable margin.

There is a significant work effort and economic savings potential in recognising that samples resulting from a proper RE(**r**) can be analysed for any number of analytes (variables). It is the same set of samples which is sent to the analytical laboratory. This constitutes a comprehensive screening of all potential analytes involved. One of these is bound to display the largest *RSV*, signifying that this analyte is exhibiting the largest heterogeneity, which means that if the entire sampling-processing-analysis procedure is focused on this

§ There has been an extensive debate at meetings and in the literature within the international sampling community as to setting up a (completely) general *RSV* threshold. Opinions have slowly converged to accept a suggestion, which originates from the characteristics of the Poisson distribution (sampling can in a certain sense be likened to a Poisson selection process), that a *RSV* larger than 20% signifies that the average of repeated sampling-and-analysis is out of control. Be aware, however, that this is an attempt to characterise all the world's extremely different materials, lots and processes with one singular threshold—a very dangerous simplification! More on this important issue can be found in DS 3077 [2] and in Pitard [1].

aspect alone, all other analytes will display an acceptably lower heterogeneity and thus cause no trouble.

Quality control of a sampling operation is completely tied in with the degree of trustworthy spatial (or temporal) "coverage" that was achieved in the deployment of the (**r**) primary sampling replications, Figure 9.3. The schematic lot depicted is meant as a *metaphor* only—it is meant to represent very many types of lots in general. For a RE(**r**) to be relevant, "coverage" is everything because this is in reality testing how an alternative singular primary sampling may come out. Having access to ten such alternative primary samples (and their corresponding analytical results) furthers critical insight into how well the particular sampling operation works in relation to the inherent heterogeneity (CH + DH) of the target lot material, i.e. how stable, or "robust" is the sampling operation in use?

There is another very useful aspect of RE(**r**). Consider that an initial sampling procedure testing resulted in a *RSV* of, say 127%. There is obviously something distinctly wrong here, 127% is way above 30% (or 20%, or lower); this represents a situation in which one would very likely find a non-representative operation somewhere in the sampling pathway. It is possible to detect precisely where the culprit part-process is to be found—for which reason the definition of full vs partial vs hierarchical replication experiments is made first.

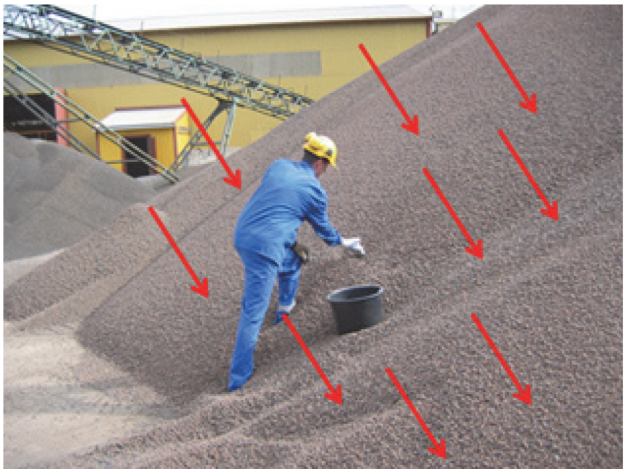

Figure 9.3: "Coverage" is everything—but not without insight. Even though the ambitious sampler is trying to "cover the ground" widely, it is also clear from a TOS point of view that the **whole** lot is far from covered properly. This illustration is a warning that deploying a RE(**10**) in-and-of itself is not enough and that it can in fact lead to misinformation, if the fundamental TOS demand of lot coverage is not properly understood.

Full replication: By the option of replication "from the top", i.e. in a baseline RE, direct, unambiguous and quantitative information as to the efficiency and validity of the *total* sampling + measurement procedure can be had. In the event of a *RSV* transgression, the message has been sounded clearly—remedying activity will be needed in order to lower the sampling quality criterion below the pertinent threshold.¶ In this fashion a full RE(**r**) delivers immediate, highly relevant information on any sampling process currently in use. The "gamble" by starting out with a full RE(**r**) is that this *may* substantiate the current procedure, in which case no further action is necessary (the gambit was won). But the opposite side of this issue is that one **must** implement relevant TOS-remedial actions in the case where *RSV* transgresses 20–30%, no exceptions! There is only one remedy for a "too high" *RSV*, be this for a full, a partial or a hierarchical Replication Experiment (**r**)—TOS to the fore for remediation.

Partial replication: It is also possible to start the replication experiment at a "lower stage" in the replication hierarchy. Figure 9.4 illustrates the general setup for full vs partial replication experiments.

Hierarchical replication: A hierarchical replication consists of the full and all partial replication setup **combined**. This is complete replication at all

¶ There exist many types of materials with significantly different heterogeneity levels. The general threshold for *RSV* (20–30%) refers to *significantly* heterogeneous materials, a very wide and diverse class of materials, for example, mineralisation, ores, pollutant effects, industrial aggregates (building materials, waste, biomass …), the list is exhaustive. There are also many other types of materials which exhibit less heterogeneity, for which a lower threshold will be relevant, say at the 10%, or 5% level, or even lower. It is emphasised that the demand for a general *RSV* is a demand which cannot be met universally. It is necessary to invoke common sense when making *RSV* operational in a distinct problem-dependent fashion [2–4].

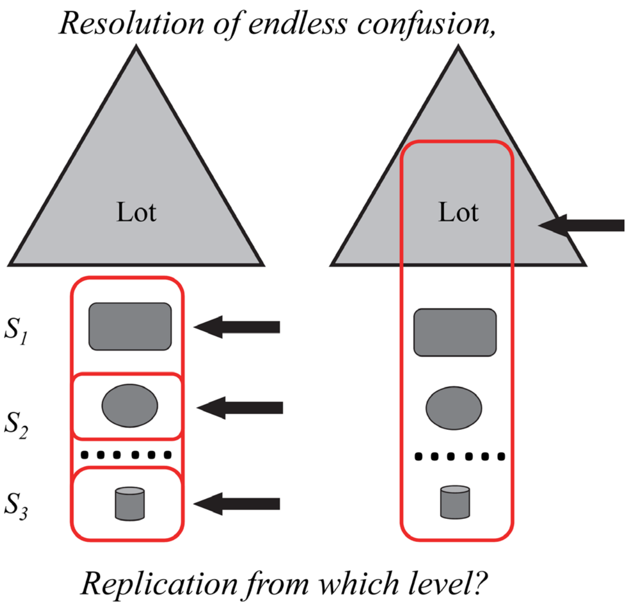
Resolution of endless confusion,
Lot
Lot
S_1
S_2
S_3
Replication from which level?

Figure 9.4: Illustrating full vs partial Replication Experiments, i.e. different starting levels for the RE (horizontal arrows). This illustration harks back to the hierarchical levels of “replication” delineated in the introduction, section 9.1.

levels (of course starting with the primary sampling). The situation is illustrated in Figure 9.5 in which the magnitude of each s^2 is represented by the length of a horizontal bar. As each lower level variance is included in any of those pertaining to higher levels, the lower level variances can be subtracted from those from all higher levels. By this simple subtraction, one can perform a complete *decomposition* of the level-specific s^2. For details, the reader is referred to http://www.spectroscopyeurope.com/sampling/sampling-quality-assessment-replication-experiment.

9.4 RE consequences for validation

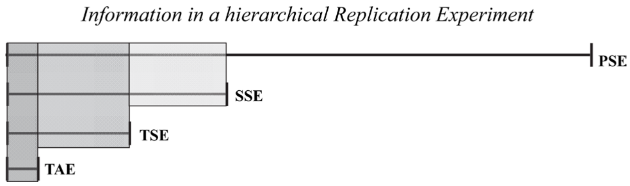

Figure 9.5: Hierarchical replication setup. A separate RE(**r**) for each sampling stage allows full decomposition of RE(**r**) variance into stage components.

From the above, it is apparent that a Replication Experiment (**r**) furnishes the exact information needed to tackle all the issues presented earlier. RE(**r**) is an unambiguous quantitative index of the total sampling-and-analysis variability induced through the full lot-to-analysis pathway. In this sense, precisely, the Replication Experiment (**r**) is a “measure of everything”. This has a direct impact on the validation issues, especially regarding the aspect of “replicate measurements”. Two easy outlines of these interrelationships can be found in Esbensen *et al.* [3–4]. The next section discusses these principles in more detail focusing on development of multivariate calibrations for spectroscopic applications.

9.5 Replication applied to analytical method development

Multivariate regression is typically carried out on samples, hopefully representative samples, that must cover the relevant range of constituent composition (or other properties) such that these samples in particular span what is expected in the *future*. Note here that analyte range is used as a simplified proxy for DH_L. Where the application requires collection of natural, technological or industrial samples, heterogeneity is a common and serious issue, particularly the “representativity of the training data set” with respect to the target lot/system. Additional questions also arise, more related to the analytical issues, such as “how many replicate scans should be performed?” In some industries, for example the pharmaceutical industry, although products are supposed to be manufactured to within tight specifications, there may still be serious issues of heterogeneity, for unit operations including mixing, where it is necessary to extract samples from powder blends for quality control purposes, see Esbensen and Romanach [5] for an example of this situation, or to extract tablets from the full populations of produced units for acceptance sampling.

It is important to keep track of the "lot-to-analysis" hierarchy of scales in order not to be confused. Assume that sampling issues levels 1–3 have been attended to in a satisfactory manner. The main remaining issue to be addressed when utilising spectroscopic methods [such as near infrared (NIR) or Raman spectroscopy] then concerns the heterogeneity of the "analytical sample". Thus, the section below deals with replication stages 4–6 only.

Spectroscopic analysis of solid materials is limited to the size of the "beam foot print" of the instrument. In the case of materials exhibiting low heterogeneity (for example, certain raw materials, or "well blended" powders), the conventional beam spot size may well be able to further a representative measure of the entire material, if and when suitably validated and verified. However, in the case of more heterogeneous materials, including natural samples of fruits, vegetables, grains, aggregates, soil … the characteristic sample heterogeneity will be much larger than the beam/spot size scale footprint (a particular issue with quantitative NIR and Raman spectroscopy).

In such cases, replication takes the form of taking a number of scans over one, or more regions of the surface of the analytical aliquot and averaging to obtain a spectrum that minimises the inherent variability encountered at this analysis stage. Many vendor solutions to this particular issue can be found, the common feature of which is to enlarge the analytical area/volume of the effective footprint by mechanical means, including the use of a rotating sample dish or a translating beam. A particularly effective way to increase the analytical area is to analyse from the outside of a rotating and translating cylinder. With such solutions, the effective analytical area can be increased 10-fold to 50-fold, allowing very powerful coverage and averaging to come into play. This approach could be termed area-enhancing spectral-acquisition replication, but it is important to keep in mind that the target here is one analytical sample only (which may, or may not be representative in itself[**]).

Replicated spectral acquisition with the same spot location amounts to nothing more than an estimate of TAE, which usually has been produced many, many times over earlier; it is a particularly nonsensical, and economically wasteful, operation to mandate replicating the analysis over and over as per routine (clearly without thinking).

These features should seem obvious, but in many calibration situations, a bulk sample is still split so that half goes to the laboratory (for reference analysis) with the other half for spectroscopic analysis, tacitly *assuming* that all 50/50 split samples will always be identical. In many cases, however, calibration models have poor precision due to this type of *disconnect* between the measured sample and the related laboratory reference values, precisely because there may still be a significant heterogeneity at the bulk sample level; it is all again a matter of the specific heterogeneity of the material. Fortunately, there is an easy solution to this problem: it should always be the **same** sample that is measured by the particular analytical instrument which is also sent for laboratory analysis. **If** one is even to begin contemplating deviation from this mandate it is comfortable to know that representative sample splitting is entirely possible (chapter 3).

Even in the situation where the reference analysis indeed is representative of the sample scanned, there is a misconception about replication in method development. The International Conference on Harmonisation (ICH) has developed guidance to industry on how to validate analytical methods. In the document entitled "Validation of Analytical Procedures: Text and Methodology" [6] the conventional principles of Repeatability and Reproducibility are defined. These are important aspects that must be considered when replacing a primary analytical method with an alternative, secondary method. Within this context,

[**] A classical case is that of analysis of protein and moisture in wheat, where a bulk sample is introduced to the instrument and is analysed many times over to produce a single averaged predicted value, essentially following the same procedure, but can this procedure be generalised? An appropriate answer would be related to how representative the reference sample is with which the spectral replication/averaging approach is to be linked in a calibration, in addition to the specific spot size issue pertaining to the instrument. There is never analysis without (some form of) preceding sampling, sub-sampling, sample preparation etc.

the precision of an analytical method is separated into three components.

Repeatability typically measures the precision of the analytical method on the same sample (stage 6, TAE) and is to be measured within a short time-period. In this case *RSV* will of course be low, otherwise the analytical method itself would be considered to include too much random variation for the precision to be acceptable. The suggested value for *RSV* in this case is expected to be below 2% (ICH).

Intermediate Precision is a measure of how well a procedure can be performed in "a short time period", assessed over factors such as days, analysts, instruments etc. It is typically performed in a non-biased way using an experimental design that includes analyst 1 performing analyses on instrument 1 on day 1. The same *sample stock* is given to analyst 2 on day 2, using either another instrument or the same instrument analyst 1 used, but completely "reset". This is done in this fashion to ensure that the samples are "true replicates". To add statistical credibility to the results obtained, analyst 1 is typically the most experienced analyst and analyst 2 the least experienced.

The results of an intermediate precision test are statistically compared using a paired *t*-test (chapter 2), in which a significant analytical bias can be detected and its magnitude can be assessed. For example, if the bias between analysts is found to be insignificant, then there is no difference in the replicate samples being measured by the two analysts and therefore, the Standard Deviation of Differences (SDD) can be used as a measure of the Standard Error of Laboratory (SEL), which is a primary statistic used to compare the precision of the primary method to the alternative method. The SEL is discussed in detail in chapter 7.

Reproducibility is a measure of how well *different* laboratories can perform the analysis developed by one laboratory and is a measure of method robustness. It is obvious that the sample stock heterogeneity is critical here, and needs to be fully evaluated (for example using a replication experiment).

These three components may nevertheless constitute a smaller part of the total prediction error, or classification rate, compared to stratification grouping of samples due to time, raw material supplier and other uncontrolled sources of variation.

Under the auspices of ICH, when it can be shown that an analytical method has no bias—accompanied by a suitable precision, this is *usually* taken to indicate that the primary sampling, or sampling regime used, is adequate. But this conclusion is **only** applicable to certain, well-specified industry sectors dealing with demonstrable low-heterogeneity materials, most notably in the pharmaceutical industry. It is essential not to fall into the trap of *illegitimate generalisation* here (see section 11.6). Where a significant heterogeneity exists in the lot material, as well as in the sample being measured, the method development scientist must take this into account up-front when developing a relevant sampling plan. The full brunt of the heterogeneity issues treated above can be present, or be present to an intermediate degree, or not at all; the point is that one usually does not know what this status is. But this is all a matter of the characteristics of the lot material interacting with a specific (good or bad) sampling method—plus TAE, most emphatically **not** pertaining to the analytical method alone.

9.6 Analytical vs sampling bias

A critical feature concerns the relationship between the analytical bias and the sampling bias. The analytical bias, a well-known concept in analytical chemistry and metrology, signifies a *systematic* deviation of *constant magnitude*, i.e. the classic statistical bias. Applying a dedicated experiment, its magnitude can always be estimated, after which it can be corrected for by a simple subtraction (lower-right panel in Figure 9.6).

In opposition to this conventional analytical bias understanding, the sampling bias is of a distinctly different nature—the sampling bias is *not* constant, but *varying*. Because the sampling process interacts with different, spatially dislocated parts of a heterogeneous lot material every time when a "replicate sample" is extracted, repeated attempts to estimate the magnitude of the sampling bias will in principle result in different dispositions of the ensemble analytical results, as illustrated in the upper-right panel of Figure 9.6. The sampling bias is *inconstant*, and can therefore never be subjected to any form of bias-correction. This is the most fundamental difference between appropriate understanding of the analytical process and the specific issues pertaining to

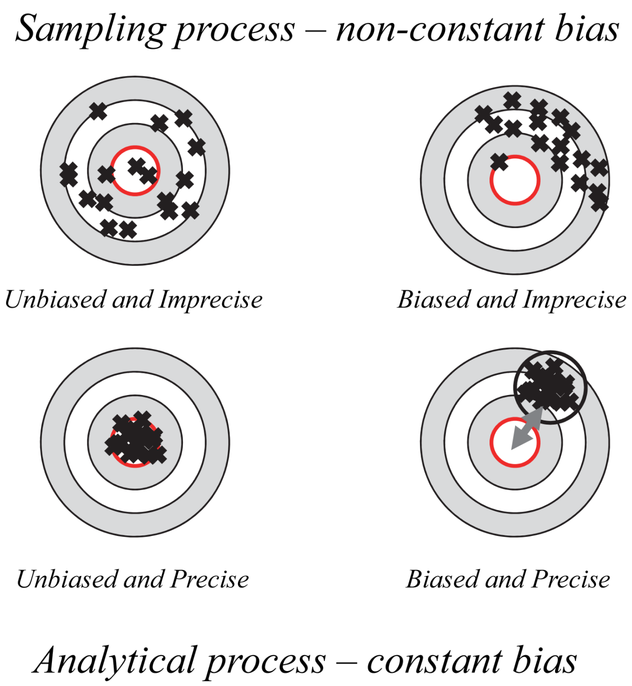

Figure 9.6: Analytical vs sampling bias. The analytical bias is per definition always assumed to be constant and can therefore be subjected to a statistical bias-correction (lower-right panel). The sampling bias is of a fundamentally different nature, however, due to **heterogeneity**. The sampling bias varies in magnitude **every time** it is estimated as a consequence of the heterogeneous nature of the lot/system and can therefore **never** be corrected for (upper-right panel). If estimated one more time the sampling variability would again be both biased and imprecise (but would constitute a *different* point swarm location and disposition in the dart board illustration metaphor used here). Instead the sampling bias has to be **eliminated**, the tools for which is to be found in TOS.

the sampling regimen.[††] A significant, regrettable confusion has existed for many decades with this root cause. These issues are also put under the validation microscope in Esbensen and Geladi [8].

TOS draws the only logical, scientific conclusion possible: the sampling bias must instead be **eliminated**. As it turns out, this is fully possible albeit with very different means than a conventional statistical correction. The salient issue is that it is impossible to know *a priori* when the case is of low, intermediate or high material heterogeneity if a heterogeneity characterisation, in the form of a replication experiment has not been performed, see above.

It is not possible to design an appropriate sampling procedure or sampling plan in the absence of information about the inherent heterogeneity met with (at whatever scale). It is a persisting myth, borne mostly out of ignorance, that sampling representativity can be achieved simply by acquisition of a particular piece, brand or type of sampling equipment or by following a standard(ised) sampling procedure—the claims of very many standards and OEMs notwithstanding. Many existing types of equipment are in fact not in compliance with the principles of TOS, and will not deliver representative samples, DS 3077 [2].

Without sampling representativity, there can be no data representativity—without which the data analyst is in reality "flying blind", delving into the complexities of multivariate calibration based on a too-limited conceptual understanding of all the types of errors influencing the total Measurement Error (MU). There is no escaping this troublesome situation, not within analytical chemistry, within statistics nor within data analysis in general. This is the reason behind chapters 3, 8 and 11 in a curriculum for data analysis.

In point of fact, there has developed a firm tradition within chemometrics surrounding multivariate calibration validation that only measurement uncertainty in the strict sense ("measurements") and the search for **the** definite validation index which can be used for all types of data has been established. The most prominent of these is undoubtedly the Ratio of Performance to Deviation (*RPD)* index used massively within the NIR community. This traditional usage has recently been put into a proper perspective by a slightly iconoclastic paper, Esbensen *et al.* [9], in which can be found a broader understanding of the straightjacket restrictions that actually pertain to *RPD*—which makes it significantly less than **the** universal validation performance indicator sought for.

[††] A full account of these interrelations can be found in Esbensen and Wagner [7].

9.7 References

[1] Pitard, F. (2009). *Pierre Gy's Theory of Sampling and C.O. Ingamell's Poisson Process Approach. Pathways to Representative Sampling and Appropriate Industrial Standards*. Doctoral thesis, Aalborg University. ISBN 978-87-7606-032-9 (available from the author at FPSC@aol.com)

[2] *DS 3077. Representative Sampling—Horizontal Standard* (2013). Danish Standards, www.ds.dk

[3] Esbensen, K.H, Geladi, P. and Larsen, A. (2013). "Mythbusters in Chemometrics: The replication Myth 1", *NIR news* **24(1),** 17–20. https://doi.org/10.1255/nirn.1390

[4] Esbensen, K.H., Geladi, P. and Larsen, A. (2013), "Mythbusters in Chemometrics, 6: The Replication Myth 2: Quantifying empirical sampling plus analysis variability", *NIR news* **24(3),** 15–19. https://doi.org/10.1255/nirn.1364

[5] Esbensen, K.H. and Romañach, R.J. (2015). "Proper sampling, total measurement uncertainty, variographic analysis & fit-for-purpose acceptance levels for pharmaceutical mixing monitoring", *TOS forum* **Issue 5,** 25–30. https://doi.org/10.1255/tosf.68

[6] "ICH Harmonized Tripartite Guideline Q2(R1), Validation of Analytical Procedures: Text and Methodology" (1997). *Federal Register* **62(96),** 27463–7.

[7] Esbensen, K.H. and Wagner, C. (2014). "Theory of Sampling (TOS) versus Measurement Uncertainty (MU) – a call for integration", *Trends Anal. Chem.* **57,** 93–106. https://doi.org/10.1016/j.trac.2014.02.007

[8] Esbensen, K.H. and Geladi, P. (2010). "Principles of Proper Validation: use and abuse of re-sampling for validation", *J. Chemometr.* **24,** 168–187. https://doi.org/10.1002/cem.1310

[9] Esbensen, K.H., Geladi, P. and Larsen, A. (2014). "The *RPD* myth…", *NIR news* **25(5),** 24–28. https://doi.org/10.1255/nirn.1462