

Validation: An invisible Component of Measurement

[Version: 2003-09-04]

Whenever we analyze a test sample, validation is lurking in the background. In fact, whenever we make a measurement or take a reading, how do you know your value is correct? Usually it is based on faith and intuition the answer seems right so there is no reason to question it. Only when a measurement result doesn't fit into an overall picture do you begin to question the performance of the measurement system. But analytical measurements, chemical and microbiological, are not self-correcting. Many operations are so common that we take them for granted. When did you ever find a miscalibrated weight or a mismarked ruler? Only because your standard operating procedures and a potential third party assessment require checking your basic measuring tools (balances, weights, pipets, burets, and volumetric flasks) do you check these fundamental analytical instruments. When an irate customer doesn't like or questions a reported value and requests a remeasurement, or sends it to another laboratory, is a discrepancy likely to be discovered.

Validation is the process of verifying or demonstrating correctness. Obtaining a value for a measurement is not sufficient; the object is to report the "true result." Discussion of the concept of "truth" has occupied philosophers for millennia. But scientists have developed their own concept of being sufficiently close to the true or correct result that permits them to operate efficiently in the real world. They set up boundaries or limits to how far results can deviate from the "true value" and still be acceptable as the "correct value." Knowing that even liberal limits can occasionally be transgressed but still be acceptable, second order boundaries are also set up. These are expressed in terms of how sure you want to be of the correctness of the result. These are often set on the basis of the consequences of making an incorrect decision. Practical analytical chemists need not become philosophers to report useful results. All they have to do is to demonstrate correctness in terms that their associates or customers understand.

There are a number of ways of demonstrating correctness of analytical results as long as we are dealing with classical stoichiometric chemistry that is taught in our classes of analytical and physical chemistry. We first learned of reactants, products, and reactions that go to completion; then solubility products and equilibrium for reactions that don't quite go to completion, but which can be forced to do so by an excess of reactant or removal of a product. But these fundamentals involve the easy inorganic analytes. There is usually a fairly pure analyte available as a basis for a parallel analysis in a comparable matrix for a demonstration of correctness. Occasionally there exists a reference material, certified by a metrological institute, of comparable complexity that can be analyzed simultaneously to provide a basis for demonstrating comparability. In geology, for example, a number of reference powdered rock materials have been available for many years and the U. S. National Institute of Standards and Technology (NIST) is now issuing standard reference materials (SRMs) for empirical analytes in food.

A problem arises when we leave the comfortable theory of stoichiometric chemistry and have to rely upon calibrated instruments to measure physical signals from our analytes, which must then be related to concentration. The problems usually occur at very low and very high concentrations, beyond the readability of our balances and burets. At these levels we have to provoke our chemical analytes into producing physical signals, or their converse, which are easily read by instruments, primarily optical and electrical, and which have a mathematically defined relationship to concentration. The most useful optical signals of light absorption in the visible and ultraviolet are transformed into a linear electrical signal, while initial electrical detector signals derived from voltage changes generated by the presence of an analyte are usually linear or logarithmic. When immunoassays generate measurable reaction products, the relationship is usually exponential.

But typically the measurements are not made on simple solutions of test samples. The analyte we seek is often buried in a sea of other compounds; some that react like our analyte and others serving merely as a diluent or carrier. We develop a procedure, often by chromatography these days, that isolates the analyte of interest from everything else and presents it in a relatively pure form to a detector that generates the measurable signal. There is generally a long chain of procedures that must be traversed if we are to find the proverbial needle of our analyte in the environmental haystack. The first and probably most important is sampling.

Sampling

Although you may not think of it as such, whenever you remove a portion of material from a larger quantity and use it to represent the entire amount—a cup of milk from a quart carton, a spoonful of flour from a retail bag, a trierfull of contents from a bag or pile, a tablet or capsule from a bottle—you are removing a sample. It is a representative sample only if the parent material is homogeneous, like the cup of milk. If the parent material is lumpy, or if it consists of different materials such as sand and gravel or peanuts and cashews, you may not get a good or representative sample, but nevertheless it is a sample of the parent material.

There is only one way to validate sampling - examine all the remaining material. The only way to be absolutely sure of the composition of a lot is to examine the entire lot, an absurd situation. Therefore you must be content with sampling and it is up to you to make a decision, uncomfortable as it may be, of how sure you want to be that your sample represents the lot. Several factors enter into the decision. The most important one is based on the consequence of being wrong. If you are examining for pathogenic bacteria, the consequence of not finding the organism is serious, but if you are sampling to see if sufficient salt has been added to your pickling brine an incorrect concentration may merely result in complaints of an off flavor. Sampling for net contents is serious from the point of view of the economic health of either the producer or the consumer and permits the employment of statisticians to develop optimal strategies to balance the risks of giving away too much of a product against cheating the customer.

Validating sampling requires resampling, balancing the consequence of being wrong with the cost of being right. The variables involved are fairly simple and easily understood: the size of the lot; the cost (number) of the samples, the labor of sampling, and the cost of

examining the samples. Incidentally, an interesting finding from the study of sampling is that increasing the size or number of samples indefinitely is a grossly inefficient use of resources. There is a point beyond which little additional information is accumulated per additional sample. The situation with respect to validation of sampling is that it is impossible to validate the correctness of sampling without examining the entire lot, an impractical situation. The matter is handled by arbitrary rules, which assign certain risks to being wrong to both the consumer and to the producer. These rules, however, only pertain to situations where economic costs can be assigned to both parties. Where a failure affects the health of the consumer, the law does not permit a balancing of risks but proof of the presence of a harmful ingredient in any sample is sufficient to trigger sanctions.

Analysis

When dealing with chemical or microbiological analysis, validation requires a demonstration that the methods used produce useful and replicable results. You can never be sure that they will produce the "true" or correct result. You have to be satisfied that they produce results of acceptable uncertainty to be useful. Analytical measurements of chemical substances (analytes) or microbiological organisms are not constant, replicable numbers, but rather the result of assigning a best estimate from one or more measurements and examining these best estimates with respect to how well they can be accepted as representing the "true" or correct result. There is a subtle interaction here between the method and the analyst, but for simplicity, we initially ascribe all the variability to the method, keeping in mind that ultimately we will have to introduce additional arenas of interest when we transfer that method to other analysts and to other laboratories.

The best estimate requires using a method that exhibits satisfactory performance with respect to a number of properties of measurements. These properties will include the following general characteristics:

- Identity (markers that distinguish an analyte from all others)
- Applicability (pertinent, relevant, suitable)
- Concentration range and limits
- Accuracy (absence of bias)
- Precision (repeatability and reproducibility)
- Selectivity (isolation of signal from others)

These properties only apply to a fully developed, optimized, stabilized, and standardized method. Methods characterized at earlier stages require revision, recharacterization, and revalidation. Validation consists of demonstrating satisfactory performance with respect to these characteristics.

Identification: The identification of a completely unknown analyte is probably one of the most difficult feats of organic and analytical chemistry. Nature has synthesized some very strange structures and organic chemists have applied the tools of infrared spectroscopy, mass spectrometry, and nuclear magnetic resonance to unravel those

configurations. Once a few milligrams of a characterized reference compound are available, validation of the identity of the compound of interest is easily accomplished by application of the principle used long ago with mixed melting points. No new peaks appear in any of the powerful separation techniques or in the energy spectra when a mixture of approximately equal amounts of the compound of interest and the reference compound are cochromatographed or examined spectroscopically, ignoring minor impurities.

So many organic compounds absorb in the ultraviolet that this property is not a very useful indicator of identity, unless a very selective isolation technique has preceded the spectral examination.

Quantitation: One of the aspects of method development that will be useful for the validation of the method is to identify the critical aspects of the method early in the formulation stage. These are the critical control points of the method that must be adhered to in order to obtain reproducible results. Some examples include the solvent, time, and temperature of extraction of active ingredients from plant materials, the requirements for purification of the extract by transfer of the active ingredient or the impurities by reextraction into a second solvent or by solid phase extraction, and the adjustment of conditions for chromatographic separations and measurement. These can be identified during method development or by application of Youden's factorial technique (1).

System suitability characteristics: Related to critical control points are the system suitability characteristics of chromatographic procedures. These are the values assigned to the variables that are required to obtain optimum chromatographic performance. The major variables are such things as the nature and size of the column, composition of solvents, the temperature and pressure of operation, and the operating characteristics of the detector. Minor deviations in these system parameters are permitted in order to maintain optimum peak characteristics and separations. Some adjustments may be required in pH, solvent composition and gradients, column temperature, and similar operational requirements. Columns deteriorate with age and changes in operating conditions are required to maintain their original separation capacity. The ability to make such changes is also important when a small tailing peak may be present along with a major peak. Minor peaks may be sought by using a diode array UV detector for peak purity calculations, the detector wavelength may be changed if the structures are quite different, or a mass spectrometer may be used as the detector. Another technique is to collect the tail of the major peak separately and reinject it; the relative concentration of a minor peak, if present, may be increased sufficiently to be displayed as a second peak.

Applicability: The most general requirement of a method of analysis is its range of application: to what products does it apply; to what products does it not apply; the approximate lowest and highest concentrations of measurable response; in the presence of what other expected ingredients; in the absence of what other common ingredients; and the availability of alternative methods. The statement must be based upon reasonable expectations and cannot be expected to encompass the universe of potential applications.

Reliability characteristics: The reliability characteristics that must be examined for validation include:

- Accuracy as the difference from the "true" or accepted value
- Precision as the difference from similar or related values repeatability (simultaneous or consecutive replicates)
- Intermediate precisions in the same laboratory by:
 - different analysts
 - different instruments
- Reproducibility (different laboratories)

For single laboratory validation, the examination of the reproducibility parameter as well as the behavior of the other parameters in different laboratories is postponed until the single laboratory demonstrates that the method has the potential for practical routine applications.

Calibration curve: Validation, or determining the reliability characteristics of a method, particularly those that depend on instruments, begins with the preparation of the calibration curve (also known as the standard curve or reference curve) that relates the concentration of the analyte to the measured signal it produces. Although it never appears as a reportable output from analytical work, it is the critical point of modern calibration or instrumental analytical chemistry. It is typically constructed by preparing a series of 5-10 analyte concentrations covering the response range of the instrument, including a blank (0 concentration) and plotting the response as the y axis against the concentration on the x-axis. If the instrument is computer controlled, it will plot the curve automatically or retain the values internally as a basis for calculation. Most instructions also require the calculation of the equivalent of a correlation coefficient, R^2 (calculated almost automatically by most spreadsheets), because most such curves, even if they have a pronounced departure from linearly, will give a value close to 1.00, which can be pointed to (incorrectly) as a satisfactory output. The curve is actually used in an inverse manner: A signal is read and is transformed by the curve to a concentration. When used in this manner the curve is called an analytical curve.

The important aspect of calibration curves is not their relationship to linearity but their stability how well they can be replicated at different times on the same day and on different days. Absolute differences between instruments can be expected. Curves prepared on one instrument ordinarily cannot be used on another unless their measurement scales have been calibrated and adjusted to the same standards.

For calibration purposes, the shape of the curve, linear or otherwise, is almost irrelevant. A linear calibration curve is convenient; it simplifies the calculations and permits use of fewer standards. But it is the replicability and stability of the calibration curve that is critical. The curve should be prepared several times, using different sources of the standard, if available, at different times on the same day, and on different days. If the curve is not stable, the method must provide for incorporating one or more standards with every batch of analyses to compensate for drift. The documentation of validation should demonstrate the stability of the calibration curve.

For a general method, applicable to a wide range of concentrations, the standards should be distributed more or less evenly over the instrument range. If trace quantities are sought, the standards should be concentrated at the lower end. Standards at the upper end must not be neglected, however, because they are needed to establish the slope more reliably than a cluster of low values. The slope is needed to establish the intercept, which in turn is needed for determining the lower limit of analysis. If a pharmaceutical preparation with a constant dosage is the object of the analysis, the standards may be arranged to bracket the expected concentration. In fact, for single tablet analyses, a single standard of the expected concentration is the usual reference for quality control work.

In special cases, where there are substantial losses in the procedure or where matrix interference contributes to the final reading, the standards may be carried through the entire procedure or they may be added to a matrix blank and the analyte is determined by the "method of addition." Other techniques include the use of an internal standard, a compound that behaves like the analyte but which can be differentiated from it, such as an isotopically labeled material. These procedures are less desirable because they are lengthy, more expensive, or less precise than a separate calibration curve because of the law of propagation of error.

Accuracy (bias): The term "accuracy" has acquired so many meanings that some organizations no longer use it, preferring to substitute terms with specific interpretations such as "bias", "trueness", or recovery. We will use accuracy to mean a difference from the "true" or accepted value, regardless of the source of the difference, and regardless of how many values and their arrangement entered into the calculation. For validation purposes we want to know if the long-term average gives us a value close to 100% recovery, if there are tolerable losses in the procedure, or if there are interferences which lead to values greater than 100%. But any specific value or even any short-term set of values is likely to be distorted by variability, so any single investigation of bias may be misleading. Like the calibration curve, accuracy should be validated over a period of time in order to average out the distortions introduced by variability.

If the bias is introduced at the signal measuring stage, any changes to the method directed to improving this step will affect both the calibration curve and the measurement equally. One hidden source of error lurks in the preparation of calibration standards from a single stock solution. If the stock solution is incorrect, all calibration solutions prepared from it are similarly in error. Such an error is only disclosed by the preparation of a second independent stock solution. Attempts at improving recovery are also likely to be effective if directed at early stages of the method, particularly the initial extraction from the matrix or carrier. This usually involves changing the solvent composition, extending the time, or elevating the boiling point of low boiling solvents. Soxhlet extraction, which constantly introduces fresh solvent, should be more effective than static extraction, but if the active ingredient(s) is bound to the carrier the bonds must be broken by chemical reactions. But steps taken to improve recovery are the province of method development rather than method validation.

In any case, acceptable bias is a function of concentration. For relatively pure isolated active ingredients, accuracy of the order required by pharmacopeial monographs should be required. For diluted dosage forms or natural product source material, the limits are

necessarily wider. Table 1 suggests acceptable validation requirements for both accuracy and precision as a function of concentration, determined from a minimum of 3 independent (different times) runs at each concentration level of interest. For example, with a botanical starting material, this validation procedure would require as a minimum the starting dried plant specimens, a commercial strength extract, and a dosage form suitable for commercial distribution. If several different plant species serve as source materials, all of them should be included. If the extracts are standardized to several different strengths, specimens of each concentration should be included, as well as each dosage form of commercial products such as tablets, capsules, and solutions. If possible include a similar plant material which does not contain the active ingredient(s) as a natural blank.

Precision: Precision is the degree of closeness of repeated values to each other. It is generally independent of recovery but it is very definitely a function of the concentration at which it was measured. It is generally measured concurrently with accuracy with one major difference. A single reading can provide an estimate of accuracy but a single reading cannot determine precision because the calculation includes a factor of $(n - 1)$ in the denominator. When n (the number of values) is 1, $(n - 1)$ is 0 and the function is indeterminate.

Determining precision in a single laboratory provides a restricted view of this reliability function, particularly when the work is performed in the laboratory of the developer of the method, who is completely familiar with its performance. Experience with method performance shows that precision estimated in the originator's laboratory consistently underestimates the precision that will be exhibited when the method is exposed to the rigors of novices and experts in multiple laboratories. In fact, the precision measured within a laboratory, designated as repeatability precision, is typically one-half to two-thirds that of the precision exhibited in many typical laboratories, the reproducibility precision.

Therefore the precision estimates obtained in a single laboratory should attempt to provide a realistic estimate of performance, not a "best" estimate. Some rules of thumb to attain this objective include:

- Use an analyst that has not been intimately involved in the development work;
- Provide a wide representation of products (n) for which the method is intended;
- Replicate the work over a period of several days (d). Perform the analyses in at least duplicate (r).

Typically for single lab validation perform r (2,3, . .) replicate analyses of m test samples over a period of d days for each product type (matrix) n .

- $r \times m$ should never be less than 10
- n should be at least 2, preferably more
- d should be at least 2, preferably more.

The calculated HORRAT value, adjusted for repeatability, should lie between 0.3- 1.3.

Table 1. Recommended recovery and precision limits for single laboratory validation:

Concentration	Repeatability (%)	Recovery (%)
100%	1	98-101
10%	1.5	95-102
1%	2	92-105
0.1%	3	90-108
0.01%	4	85-110
10 μg/g(10 ppm)	6	80-115
1 μg/g(1 ppm)	8	75-120
10 ng/g(10 ppb)	15	70-125

The HORRAT value is the result of an empirical summary of the examination of over 10 000 individual interlaboratory tests conducted over the past century on materials that included foods, drugs, feeds, fertilizers, pesticides formulations and residues, drug and contaminant residues in foods and tissues, waters, rocks, minerals, and certified reference materials (2, 3). The data could be summarized by a simple exponential equation relating concentration, C, expressed as a mass fraction (e.g., 1% = 0.01; 1 ug/g = 0.000 001) to *among-laboratories* standard deviation (S_R) or relative standard deviation (RSD_R) as follows:

$$(1) \quad RSD_R = 2C^{-0.15}, \text{ or equivalently } S_R = 0.02C^{0.85}$$

$$RSD_R = S_R \times 100/C .$$

The HORRAT value is calculated by taking the actual value of RSD_R calculated from the data and divide it by the RSD_R calculated from the formula, thus,

$$HORRAT_R = RSD_{R \text{ found}} / RSD_{R \text{ calculated}}.$$

Values between 0.5 and 2.0 are considered acceptable for among-laboratory precision, expressed as a $HORRAT_R$ value. The subscript "R" is added to distinguish this value from the corresponding single laboratory value "r"., where the corresponding acceptable limits are $HORRAT_R = 0.3-1.3$.

The most interesting feature of this value is that it is more or less independent of analyte, method, matrix, and time— i.e., modern instrumental methods are no more precise than the original volumetric and gravimetric method of the last century. The precision of the Kjeldahl method is the same today as when it was introduced over 100 years ago. Of interest to single laboratory validation, within-laboratory precision, RSD_R , is approximately one-half to two-thirds of the among-laboratory precision, RSD_R . On this basis, acceptable single laboratory precision, calculated as a $HORRAT_R$ (or the more easily read $HORRAT(r)$) value would be 0.3-1.3. The data supporting the use of the HORRAT value are given in reference (2) and summarized and extended in reference (3).

Limits

Most analytical work is conducted within a reasonable working range of the method but some analytical work stretches the measurement capabilities of modern instrumentation. In fact, much of the capability of modern instruments to operate in the nanogram, picogram, and even femtogram regions was stimulated by the requirements of new environmental laws limiting residues in food, air, water, and soil. The definitions appear to be very simple: The limit of determination is the smallest amount of an analyte that can be measured reliably; the limit of detection is the smallest amount of an analyte that can be reported as present with reasonable certainty. But these definitions merely shift the problem to the meaning of reliably and reasonable. Then the problem becomes statistical in nature and the solution depends on assigning a value to the risk of being wrong. Two practical solutions have emerged:

- The first depends upon determining the magnitude of the blank, baseline, or noise when no analyte is present. Arbitrarily, the limit of detection is assigned that value that corresponds to 3 times the magnitude of the blank transformed to a concentration through the calibration curve. The limit of determination is assigned that value that corresponds to 10 times the magnitude of the blank transformed to a concentration through the calibration curve. The problem with this method is that it is independent of the analyte.
- The second solution places a value on the relative standard deviation when the method "goes out of control" and begins to produce false negatives (the analyte is present but no measurable signal is produced). This point may be deduced statistically by considering that 3 standard deviations on both sides of a mean should include 99.7% of all the observations. If we position such a (normal) distribution on the concentration axis so that the point on the concentration axis where the 99.7% point falls is "0" concentration, the mean will fall about 3 standard deviations on the positive x axis, or

$$3s = 100 (\%);$$

therefore, the standard deviation at this point is 33%. If this (relative) standard deviation is transformed to a concentration through equation (1), C is approximately 10 ng/g (ppb). This is the region of analysis for mycotoxins, and drug and pesticide residues in tissues. Below about 10 ng/g, false negatives begin to appear. This proposal assumes that the values near zero are normally distributed, an unlikely situation. These values were derived for the among-laboratories situation. A generous assignment for a single laboratory RSD_R limit situation would approximate 25%.

If a similar analysis is applied to obtaining a concentration value for the limit of determination by assigning the commonly proposed value for RSD_R of 10%, C is approximately 10 ug/g (ppm). The corresponding equivalent RSD_r of 20% would suggest that the limit of determination for single laboratory validation is in the ug/g region (ppm). Modern instrumentation is able to reach several orders of magnitude lower than these values.

Summary

Single laboratory validation is the process of demonstrating correctness with respect to identity, applicability, and reliability of an analytical method in that laboratory. The requirements for these attributes are:

Identity: Production of a unique spectrum by infrared spectroscopy, mass spectrometry, and nuclear magnetic resonance which matches that of the reference compound; produces no new peaks when cochromatographed with the reference compound.

Applicability: A statement based upon actual tests:

- to what products the method applies;
- to what products the method does not apply;
- The approximate lowest and highest concentrations of acceptable measurable response;
- in the presence of what other common ingredients;
- in the absence of what other ingredients;
- the availability of alternative methods.

Reliability: Data from known or synthetic test samples showing:

- Repeatability of the calibration curve over at least 2 days;
- Accuracy as the recovery of added quantities as a function of concentration (see Table 1); a series of at least 5 replicates conducted at least on 2 different days at least at 2 different concentrations
- Precision as the relative standard deviation of the series conducted for accuracy (see Table 1)
- Limits (if needed): of determination:
 - Concentration at which $RSD_R = 25\%$ of detection:
 - Concentration at which $RSD_R = 33\%$.

-END-