CFW AACC INTERNATIONAL REPORT

Collaborative Studies for Quantitative Chemical Analytical Methods

TERRY C. NELSEN[1] AND PAUL WEHLING[2]

A chemical analytical method is a series of ordered steps carried out in a properly equipped laboratory to estimate the concentration of a specific analyte or a physical property of a given material in an accurate and precise manner. Analytical procedures must be reliable and accurate for scientific, trade, and quality control purposes. Commodities are valued and traded based on levels of their constituents (i.e., protein, starches, oils, etc.), physical properties (color, pH, fiber, rheological measures, etc.), or biological properties (mycotoxins, bacteria, etc.). Measurements can be evaluated and compared internationally only if methods are standardized. Quality control cannot be achieved if unexplained noise exists in measurements.

Collaborative Studies

In this paper, we will discuss the use of interlaboratory studies for the evaluation of analytical methods for concentrations. Our objective is to describe procedures to validate analytical methods used for estimating concentrations of a specific analyte in a given material with a minimum of uncertainty. Many of the testing and evaluation procedures discussed here are also valid for measuring physical properties. The purpose of a collaborative study is to evaluate how a method will perform in the future in actual practical applications in real life settings. We assume that the laboratories included in the study are competent and fully capable. The collaborative study is a test of the method, not the labs. We assume that there is a true value (actual concentration) we are trying to estimate. The difference between our measurement result and the true value is the uncertainty in the measurement. This uncertainty is called the error in the method and can be divided into random error and systematic error. Random error is both over and under estimates of the true value and is usually considered to be the noise in the method. We often have difficulty in finding the causes of random error. Systematic error is a consistent bias in the system and can be eliminated or controlled if the cause can be found. We assume that the method developer has checked the method for selectivity, specificity, and linearity.

When developing a new method, the developers will often run a single lab validation (SLV) to initially test the method. After the method has been thoroughly tested during the development phase, a designed SLV can be run to demonstrate local performance. When the developers are satisfied with local tests of the method, a formal collaborative study is set up according to international validation standards. The multilab collaborative study is designed to provide confidence in the performance statistics, to demonstrate the reproducibility of the method, and to provide opportunities for improvements of the method.

A set of guidelines for standard symbols, terminology, and procedures was established at the IUPAC Workshop on the Harmonization of Collaborative Analytical Studies (2,6) and is available in the *Official Methods of Analysis of AOAC International* (3). These harmonization guidelines establish the design of collaborative studies to adequately estimate repeatability and reproducibility. The material (sometimes called the matrix) is the medium that contains the analyte. A minimum of five materials are required. The materials used as unknowns in the collaborative study should represent common matrices that the method is likely to be used on in practice. A method can be quite specific like AACC Intl. Official Method 08-12, Ash in Farina and Semolina, or more general like AOAC Intl. Official Method 991.42, Insoluble Dietary Fiber in Foods and Food Products. Method 991.42 was tested in 22 different materials and the performance results are listed for each (3).

The collaborative study requires at least eight laboratories for statistical validation of the performance parameters. We suggest that you start with 12 labs. Leave room for error, nonparticipation, or unforeseen difficulties. The labs should be representative of the labs where you expect your method will be used. We also suggest that you first run an unofficial mini-collab with two or three friendly labs to see if the method has any readily identifiable and fixable problems that had not been observed in your own lab during the development of the method and the SLV, and to ensure that the samples are homogeneous and stable. The data from this mini-collab can be included in the analysis of the whole collaborative study if no changes or improvements have been made.

We do not recommend conducting a collaborative study with an unoptimized method. An unsuccessful study wastes time and creates ill will. This caveat applies especially to methods that are formulated by committee and have not been tried in practice (3).

Decide on the concentrations where the method will be used. The collab study must test the full range of the method's scope as written. Prepare samples of analyte at levels to bracket and cover this area of interest. For example, if the method is to be used to estimate protein concentration in flours from 10 to 16% protein, then the collab study must be run with flours from 10 to 16% protein. If zero is in the area of interest, prepare blanks and consider a series of tests to determine the levels of detection and quantitation. Measuring near zero has some special considerations for statisticians (4) and will be considered in a different

[1] Member of Check Sample Committee and Statistics Committee, AACC International. USDA-ARS, 1815 N University St., Peoria, IL, 61604; Terry.Nelsen@ARS.USDA.GOV.

[2] Vice-chair of the Approved Methods Committee and member of Statistics Committee. 330 University Ave SE, Minneapolis, MN 55418; paul.wehling@genmills.com.

doi:10.1094/CFW-53-5-0285

paper. Also, this paper discusses procedures for quantitative estimates of concentrations; different procedures are to be used for qualitative (i.e., present/absent, too high/ok) testing. Blanks used for calibration or for practice runs are not considered as one of the five materials. Materials with naturally occurring concentrations are preferred over spiked samples whenever possible. Materials must be homogeneous and stable. Nonhomogeneity can cause outliers and will increase the variance estimates.

Code the test samples so it is not obvious which samples are blind duplicates and so they will not be analyzed in a set order. Prepare sets of blind or matched duplicates (a pair is considered one material). You don't really need to run triplicates; duplicates are sufficient to estimate internal variance. Rather than designing a study with five triplicates, use those same resources to design the study with seven or eight duplicates. Design a data reporting form that you send along with the samples. You want the data to come back to you in the same format from each lab.

Calculations

Calculations required for method validation are outlier tests and then performance statistics. The performance parameter estimations can be done with analysis of variance (ANOVA) procedures or by using an Excel spreadsheet that is available from AOAC Intl., which calculates both outlier and performance statistics (1). The Cochran test is used to check for outliers in individual measurements and the Grubbs test to check for laboratory outliers (8). An excessive number of outliers is more than two-ninths of the data. A method which produces outliers at greater than the two to nine ratio is considered to be unstable. An outlier can be the result of a simple mistake, but it can sometimes be the result of an unusual chemical reaction or condition. Outliers observed in the course of the collab must be investigated.

Performance parameters to be estimated are: mean or average value for each material, standard deviations for repeatability and reproducibility, relative standard deviations, repeatability and reproducibility values, and where appropriate, a HorRat. Estimates from individual materials can be combined for an overall evaluation of the method. The analysis of the data can be also done with ANOVA procedures (5,8) or with the AOAC Intl. provided spreadsheet (1).

The primary calculations are the repeatability standard deviation (s_r) and reproducibility standard deviation (s_R). The repeatability standard deviation is the within laboratory standard deviation. Reproducibility includes both the within and between laboratory variances. The between laboratory variance is not reported but is used to calculate reproducibility. In accordance with ISO Standard 5725, a repeatability value, r, is calculated as $r = 2.8 \times s_r$ and reproducibility value, R, is calculated as $R = 2.8 \times s_R$.

The precision of a method is a measure of the extent to which individual tests of the same concentration in the same material agree. Repeatability, r, is the internal precision of a method. Two single results obtained within a laboratory under repeatability conditions (same technician with the same instruments in the same laboratory at the same time) should not differ by more than r. Reproducibility, R, is the external precision of a method. Two single results obtained by two different laboratories under reproducibility conditions (different technicians with different instruments in different laboratories at different times) should not differ by more than R.

The relative standard deviation (RSD) is a useful statistic to compare different methods. The RSD is the standard deviation expressed as a percentage of the mean. An RSD is the same statistic as the coefficient of determination or CD. The RSD is independent of units of measure or of scale and becomes especially useful for experienced analysts. For a slightly different example, say we weighed 100 humans and found an average weight of 176 lbs and a standard deviation of 25 lbs. The RSD would be 15%. If we had weighed those same humans on a metric scale we would have found an average of 75.8 kg with a standard deviation of 11.4 kg and the same RSD of 15%. In our experience, we have found that mammals of the same species and variety or breed at a similar age should have an RSD of around 15%. If we weighed 100 children of the same age and found an average of 40 lbs we would expect the standard deviation of those weights to be around 15% of 40 or around 6 lbs. If the RSD is smaller or larger then we probably do not have a random sample of children of a given race and age. In chemical measurements it is common for the RSD to increase as concentration approaches zero.

At some point the analyst often asks the statistician if the r and R values are reasonable. The statistician's role in a collaborative study is to ensure that the performance parameters have been calculated correctly and not to pass judgment on the value of a chemical procedure.

A group of statisticians from the U.S. Food and Drug Administration led by William Horwitz (7) took the results of several thousand method evaluations and found a general relationship between concentration and RSD_R. They recommend that you calculate the actual RSD_R from your data and then calculate a predicted RSD_R based on the mean concentration. Divide the RSD_R from your data by the predicted RSD_R to get a HorRat value. The HorRat value should be between .5 and 2.0. If the value is less than .5, the study results are suspected of being too good to be true. A HorRat value between 1.5 and 2.0 can be an indication of material instability, among other possible problems. A HorRat value greater than 2 can cause the method to be judged unreliable and thus unacceptable. The RSD_r is usually one-half to two-thirds of the RSD_R. The AOAC Intl. spreadsheet also calculates the HorRat for you.

Table I. Percent protein of 10 wheat flour samples as analyzed by method Q1

	Flour									
	A		B		C		D		E	
	Sample number									
Lab	9	4	6	2	3	1	8	10	5	7
1	10.46	10.69	11.46	11.83	12.84	12.31	13.49	13.90	15.51	15.16
2	10.07	10.37	10.21	10.00	12.26	12.37	14.16	13.94	15.26	15.00
3	9.51	9.07	10.61	10.44	12.15	11.91	13.38	13.51	14.43	13.94
4	10.12	9.73	11.42	11.37	12.22	12.64	14.04	14.01	15.45	15.30
5	9.42	9.66	10.72	10.31	11.46	11.76	12.85	12.68	14.57	15.30
6	9.34	9.15	10.13	9.71	11.15	11.43	12.59	12.80	13.83	15.02
7	9.39	9.24	10.17	9.71	11.18	11.51	12.77	12.45	13.96	13.61
8	11.00	10.96	12.03	11.79	12.38	12.10	14.20	14.67	15.45	15.65
9	9.82	9.56	11.35	11.19	11.72	11.79	12.72	12.29	13.96	13.99
10	10.04	10.51	11.32	11.48	12.17	11.97	12.83	12.96	14.25	14.23
11	10.52	10.53	10.89	11.27	12.14	11.94	14.18	14.43	15.6	15.64
12	9.88	9.81	11.52	11.89	12.17	12.08	13.94	14.39	15.66	15.75
13	9.59	9.14	10.18	9.71	11.32	11.77	13.58	14.02	14.10	14.10
14	9.55	9.41	10.30	10.63	12.12	12.05	13.04	13.49	15.10	14.93
15	10.96	10.65	11.84	11.81	12.72	12.66	14.08	14.43	15.75	15.31

Table II. Outlier tests

Flour	Cochran test (%)	Single Grubbs (%)	Double Grubbs (%)
A	18.1	9.6	19.0
B	11.9	3.9	9.2
C	23.7	8.3	19.0
D	13.7	5.4	11.5
E	50.6	6.9	13.0
Critical value	51.5	30.0	40.9

Note that the HorRat is useful only for concentrations and not physical or biological properties. Also, the predicted RSD_R increases rapidly as the concentration approaches zero. At a concentration of 1%, you should expect an RSD_R of around 4%. At 1 ppm the predicted RSD_R will rise to 16% and at 1 ppb the predicted RSD_R will be 45%.

Accuracy

Accuracy is the extent to which the test results differ consistently from the true value. It is not uncommon for one lab to consistently overestimate a concentration in a material and a second lab to consistently underestimate the concentration in the same material. A lab can be precise but not accurate. The opposite of accurate is biased.

Consider a ranks test (5) to examine accuracy. For each sample, list the data from all of the labs and sort from largest to smallest. The largest value is assigned 1, the second largest 2, and so on. Do this for each sample. If some labs consistently have the largest or smallest values, then a bias may exist in the method. A ranks-sum test can be run to see if a significant ranking bias exists. A significant ranks test can be the result of sample deterioration if the materials were not stable and labs that ran the tests earlier reported higher values. Significant ranks tests can also be an indication on the methods sensitivity to different reagents or solvents or to laboratory temperature or humidity differences.

Example Collaborative Study

A new method (Q1) of quickly measuring wheat flour protein concentrations was developed by a team of chemists at GESR Laboratories. After preliminary tests and adjustments, they submitted the method for a collaborative study. Five different wheat flours with different protein levels were prepared. Fifteen different laboratories from around the world were contacted and agreed to take part in the study. Each laboratory received 10 flour samples that were the five flours prepared in duplicate and assigned a number by a random process. Each laboratory was asked to analyze each sample by the Q1 method and report the results estimated to two decimal places. The results received are in Table I.

These results were entered one flour at a time into the AOAC Intl. spreadsheet calculator. The spreadsheet calculates outlier tests (Table II) and performance statistics (Table III).

For this example, no outliers were found (Table II). The Cochran test looks at the difference between the estimates of the blind duplicates. It calculates a distribution of all the differences and then tests each individual difference to see if it fits in that distribution (homogeneity of variance). The Grubbs single test looks to see if the averaged value from a set of duplicates fits in the normal distribution and the double Grubbs test looks to see if any two extreme values don't fit in the distribution.

In this example, none of the tests picked up any outliers. To identify an outlier in the Cochran test, the calculated statistic would have to be larger than 51.5% for 15 labs and 2 replicates. The single Grubbs statistic would have to be larger than 30.0% and the double Grubbs statistic would have to be larger than 40.9% for us to consider a value to be an outlier. These critical values for the Grubbs tests are expressed as the percent reduction in the standard deviation caused by removal of the suspect value(s) (see reference 3, Appendix D). If any values were found to be outliers, we can simply eliminate them and rerun the analysis without them. (We encourage you, however, to look closely at any outlier and to try to determine what caused it. It may have been a simple mistake, like a misplaced decimal or a mislabeled test tube or simply writing the numbers wrong, in which cases simply discard that number. In some cases the outlier may have been a true value and was a result of a special set of circumstances in the method. Be sure that this is not the case before you discard that outlier.) You can discard up to 22% (2/9) of your data and still get valid results. You can also treat missing values in the same way—analyze without them as long as they are not more than 22% of your data.

For performance evaluation, the spreadsheet program calculates for each flour: the mean protein, s_r, s_R, RSD_r, RSD_R, r, R, and the HorRat value. Table III contains a summary of the results for each flour.

First, look over your results to see that none of them stand out. Is there one flour that is different from the rest? Is the method less reliable in one class or type of wheat flour? Are the RSD estimates related to the mean? You can run a correlation analysis but for only 5 flours the correlation coefficient would have to be greater than .81 to be significant at $P < .05$. In a test like this, construct a plot or scatter diagram of the suspected relationship and the committee will have to make a judgment call.

The HorRat values in this example range from 1.35 for flour C to above 2.00 in flours A and B. A HorRat above 2.00 is indicative of an unreliable method. Also, remember that the RSD_r should be one-half to two-thirds of the RSD_R value. In this example, the RSD_r is less than one-half of the RSD_R value, another indication of unreliability.

At this point, the committee may reject the method as unreliable for reproducibility purposes. The method developer can

Table III. Method performance statistics

Flour	Mean	s_r	s_R	RSD_r	RSD_R	r	R	HorRat
A	9.94	0.20	0.60	2.03	6.09	0.56	1.69	2.15
B	10.91	0.25	0.76	2.28	6.92	0.70	2.12	2.48
C	12.01	0.20	0.45	1.66	3.73	0.56	1.25	1.35
D	13.53	0.23	0.71	1.72	5.23	0.65	1.98	1.93
E	14.86	0.31	0.71	2.05	4.76	0.86	1.98	1.79

s_r—repeatability standard deviation; s_R—reproducibility standard deviation; RSD_r—repeatability relative standard deviation; RSD_R—reproducibility relative standard deviation; r—repeatability value; R—reproducibility value.

Table IV. Protein values ranked by laboratory

	Flour										
	A		B		C		D		E		
	Sample number										
Lab	9	4	6	2	3	1	8	10	5	7	Sum
1	4	2	4	2	1	4	8	8	4	7	44
2	6	6	12	12	4	3	3	7	7	9	69
3	12	15	10	10	8	10	9	9	10	14	107
4	5	8	5	6	5	2	5	6	5	5	52
5	13	9	9	11	12	13	11	13	9	5	105
6	15	13	15	13	15	15	15	12	15	8	136
7	14	12	14	13	14	14	13	14	13	15	136
8	1	1	1	4	3	5	1	1	5	2	24
9	9	10	6	8	11	11	14	15	13	13	110
10	7	5	7	5	6	8	12	11	11	11	83
11	3	4	8	7	9	9	2	2	3	3	50
12	8	7	3	1	6	6	6	4	2	1	44
13	10	14	13	13	13	12	7	5	12	12	111
14	11	11	11	9	10	7	10	10	8	10	97
15	2	3	2	3	2	1	4	2	1	4	24

search for the cause of the unreliability and try to fix it. One tool the analyst can use is a ranks test.

Example Ranks Test

First rank the protein values for each sample from largest to smallest. Replace the protein values in Table I with the rank value for each (Table IV). Then add up the 10 ranks for each lab. That sum is the rank sum and can be tested for significance. If the method is truly independent, then no lab should be consistently higher or lower than the other labs. For 15 labs and 10 samples, no rank sum should be ($P < .05$) less than 41 or greater than 119 (8).

In our example, labs 8 and 15 are consistently low in their estimates and lab 6 and 7 are consistently high. This result shows that the method has a systematic bias. The method may not be invalid if this bias is judged by the Methods Committee to be trivial. The method developer should, however, try to find the source of this bias and correct it.

In conclusion, if you wish to have your method officially accepted, we advise you to get a statistician before you start and have that person review the appropriate literature. Check with the Approved Methods Committee or equivalent in the organization where you want your method listed. Design the study and run an SLV and a prelim/pilot study. Finally, as always, look at your data to aid in interpretations.

References

1. AOAC International. AOAC International interlaboratory study workbook blind (unpaired) replicates, version 2.0. Published online at www.aoac.org/stats/AOAC_BlindDup_v2-0.xls. The Association, Gaithersburg, MD, 2006.
2. AOAC International. Guidelines for collaborative study procedure to validate characteristics of a method of analysis. Revised—AOAC Official Methods Program J. AOAC Int. 78(5):143A-160A, 1995.
3. AOAC International. *Official Methods of Analysis of AOAC International*, eighteenth edition. The Association, Gaithersburg, MD, 2005.
4. Currie, L. A. Limits for qualitative detection and quantitative determination: Application to radiochemistry. Anal. Chem. 40:586-593, 1968.
5. Delwiche, S. R., Palmquist, D. E., and Lynch, J. M. Collaborative studies for cereals analysis. Cereal Foods World. 50(1):9-17, 2005.
6. Horowitz, W. Protocol for the design and interpretation of method-performance studies: Revised 1994. Pure Appl. Chem. 67:331-343, 1995.
7. Horwitz, W., and Albert, R. The Horwitz ratio (HorRat): A useful index of method performance with respect to precision. J. AOAC Int. 89(4):1095-1109, 2006.
8. Wernimont, G. T. *Use of Statistics to Develop and Evaluate Analytical Methods.* W. Spendley, ed. AOAC International, Gaithersburg, MD, 1985.