Collaborative Studies for Cereals Analysis[1,2]

S. R. Delwiche
USDA/ARS, Beltsville Agricultural Research Center
Beltsville, MD

D. E. Palmquist
USDA/ARS, National Center for Agricultural Utilization Research
Peoria, IL

J. M. Lynch
Cornell University
Ithaca, NY

Analytical laboratories, including those specializing in cereals analysis, routinely apply quantitative tests whose levels of precision are traced to documented experiments that form the basis of each test. Such experiments are typically known as collaborative studies, interlaboratory studies, or "ring tests." The primary purpose of a collaborative study is to document and demonstrate the precision and, to a lesser extent, the accuracy of a chemical procedure that measures the concentration or continuum response of an ingredient, element, chemical compound, toxin, or property. Under the direction of the AACC Approved Methods Committee, the association has an established policy (for more than 20 years) of offering a method approved status after the method has undergone the rigor of a collaborative study. Approved methods are the backbone of AACC's analytical program and can be purchased from the association in either print or electronic format (1).

Briefly, this article is designed to provide a short primer on the nature of the collaborative study. AACC has generally adopted the guidelines of AOAC INTERNATIONAL (AOACI) (2). Both societies signed a memorandum of understanding several years ago on the sharing of each other's methods. AOACI's guidelines stem from an initial workshop of the International Union of Physical and Applied Chemists (IUPAC) held in Geneva, Switzerland, in 1987 (5). The guidelines have since been updated at least twice by IUPAC during the 1990s. Greater detail on the design and execution of collaborative studies can be found on AOACI's website (www.aoac.org) (3).

We begin by describing the basic criteria for a collaborative study and then explain the statistical definitions for the two terms (repeatability and reproducibility) that define the precision of a method and provide examples of how these are applied to cereals tests. We continue with a discussion of the generalization of the statistical procedure in terms of analysis of variance (ANOVA) and user-friendly software that is currently available at no charge to the analyst.

Collaborative Studies

A collaborative study consists of a number of participating analytical laboratories that each follow a defined set of instructions on the analysis of a common set of laboratory samples. The work is coordinated by, using the parlance of AOACI, the "study director" (Fig. 1). At a minimum, eight valid collaborating laboratories performing quantitative chemical analysis are needed to maintain a reasonable statistical confidence interval for reporting the precision of a method. It is generally recommended that several more laboratories be included than the minimum number as a guard against unforeseen circumstances, such as laboratories that are unresponsive, that fail to rigidly follow written directions, or that make identifiable mistakes in the procedure. Collaborators are defined as unique in both person and physical location.

As a collaborating laboratory, a minimum of five different "materials" are analyzed for the constituent or property of interest in accordance with the set of instructions from the study director. The definition of "different materials" is somewhat arbitrary. The purpose of requiring a minimum of five different materials is to instill a degree of robustness in the method. Often, different materials can be different major ingredients in a formulation, such as wheat, barley, or corn in mixed meals. In a narrower sense, the materials may in fact be more botanically similar, such as a method for ash content that is applicable to hard and soft classes of wheat.

Keys to a Successful Collaborative Study

One of the unique aspects of a collaborative study is the number of scientists and laboratories that become involved and have a stake in the study's success. Prospective study directors are strongly advised to carefully design procedures that have a high probability of success, to avoid the incidence of ill will, which can condemn the study, and, even more problematic, the notion of collaborative studies in general. The characteristics of a successful collaborative study can be summarized in six items (Sidebar).

Traditionally, a collaborative study involving a particular procedure has been performed only after years of experience by at least one laboratory (preferably the study director's), during which time the ruggedness of the procedure has been examined. The one-laboratory-leader approach is rapidly being replaced, however by a consensus approach, in which interested parties review all applicable methods and choose the best one.

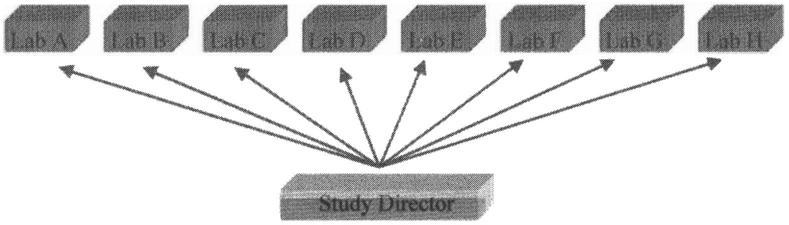


Fig. 1. Collaborative study design.

[1] Originally presented (abstracts 121 and 122) at the 2003 AACC Annual Meeting, Portland, OR.

[2] Mention of trade names or commercial products is solely for the purpose of providing specific information and does not imply endorsement or recommendation by the USDA.

Ruggedness of Procedure. By "ruggedness," we mean that the procedure has been examined for the effect of the variation of typical experimental conditions that exist in analytical laboratories, such as those caused by different sources of reagents, equipment, temperature, pressure, and humidity. A rugged procedure is one that has a tolerance for the variation in conditions that can be expected from laboratory to laboratory.

Writing a Protocol. Preliminary testing may include an examination of the specificity of the procedure and how it is affected by contaminants, additives, and trace ingredients that are bound to occur in commercial process settings. A carefully written protocol, which consists of the method instructions for collaborators and the complete study design for the committee (such as the AACC Approved Methods Committee) overseeing the study is essential to the study's success. The method instructions should be a stand-alone set of directions, complete with specification of glassware sizes, buffer solution preparation, calculation constants, and any other resources that an experienced analyst would otherwise need to complete the procedure.

Pilot Study. Because of the size of the undertaking, collaborative studies are often preceded by a pilot study, in which a few laboratories are asked to perform the procedure. These laboratories may become collaborators in the ensuing study, although the results of the pilot study are not to be used in the latter. The purpose of the pilot study is to uncover unforeseen sources of variation in the procedure and to reveal potential misunderstandings in the method instructions—all with the intention of remedying these before beginning the actual study.

Stable Materials. Realistically, collaborative studies are performed over the course of several weeks, as collaborating laboratories, which are often commercial analytical laboratories, adjust their workload to accommodate the additional assays and possible new analytical procedures. Therefore, the stability of additives, reagents, and laboratory samples must account for this extended time period, with advisement of storage conditions of such materials noted in a letter of transmittal included with the laboratory samples.

Ingredients for a Successful Collaborative Study

- Rugged analytical procedure
- Clearly written protocol
- First tested in a pilot study
- Stable materials
- Test (sub-)samples representative of the whole
- Dedicated collaborators

Sample Uniformity. To ensure that each collaborator's laboratory samples are the same as those of all other collaborators, careful attention must be given to their assembly. In cereals studies, the materials are usually collected first in bulk, which in the terminology of IUPAC (6) we will refer to as an "aggregate sample," then subdivided, or split, into the requisite size for each laboratory (i.e., "laboratory sample"). Combining of ingredients and additives preferably is done at the bulk level. However, this assumes that a suitable means of mixing to a uniform consistency is possible. Grain dividers should be used if the materials are in their native state. Because it is the method and not the materials themselves that are under evaluation, intense effort should be put toward uniformity in the laboratory samples sent to each laboratory. A representative number of laboratory samples should be taken at random and tested for uniformity before distribution to collaborators. Further attention is required of the collaborator when the received laboratory sample is reduced to the size needed for actual analysis, termed the "test sample," which is termed the "test position" during later analytical operations. Otherwise, unknown levels of variation will affect the method's precision.

Collaborators. Participation in a collaborative study should come about through the least level of coercion and with the full support of the laboratory management. A willingness to participate in the study is a good indication that the collaborator will follow directions and return the requested data. Relying on a few disinterested collaborators who fail to either follow directions or submit data may place the entire study in jeopardy due to the failure to meet the eight-laboratory minimum.

Method Precision

The two terms that have well-defined statistical meanings when discussing a method's precision are "repeatability" and "reproducibility." By emphasizing these two terms, an underlying implication is made that although the accuracy of the method, which can be defined as its closeness in values to those of some "gold standard," may be of importance, the accuracy of the method is of less concern than the method's ability to be consistent, both within the same laboratory and among different laboratories. Repeatability characterizes within-laboratory precision, while reproducibility encompasses this source of variability, and, more importantly, also includes the variability associated with method determinations from multiple laboratories.

Repeatability

—"A measure of how well an analyst in a given laboratory can check himself using the same analytical method to analyze the same test sample at the same time" (2).

Reproducibility

—"A measure of how well an analyst in one laboratory can check the results of another analyst in another laboratory using the same analytical method to analyze the same test sample at the same or different time" (2).

Most often, repeatability is determined from the analysis of "blind replicates." As implied by the name, blind replicates are laboratory samples, usually two and seldom more than three, for which the collaborator is unaware of their identical nature (i.e., they are from a common aggregate sample). Acknowledging that this may not be true, collaborators may recognize the similarity of two laboratory samples by appearance or closeness in measured value, then, without malice, perform actions to heighten the closeness in their values. To circumvent this potential problem, the concept of using closely matched, but not truly identical, pairs, collectively known as "Youden matched-pairs," is an alternative approach (discussed later).

Assuming for the moment a study in which blind duplicates are used, repeatability and reproducibility can be evaluated from simple statistical calculations, with the conventional assumption that repeated measurements on a laboratory sample are normally distributed. The repeatability standard deviation (s_r), as evaluated on a blindly duplicated laboratory sample with a difference between duplicates of d_i for each of n laboratories, becomes

$$s_r = \sqrt{\sum_{i=1}^{n} d_i^2 / 2n} \qquad (1)$$

Likewise, the reproducibility standard deviation (s_R) will be

$$s_R = \sqrt{(s_d^2 + s_r^2)/2} \qquad (2)$$

The first term within the parentheses of the above expression is determined as

$$s_d^2 = \frac{\sum_{i=1}^{n}(T_i - \overline{T})^2}{2(n-1)} \qquad (3)$$

such that T_i represents the sum of the individual values within laboratory i and $\overline{T}$ is the mean over the n laboratories. In its more general form, in which more than one material is used to characterize precision collectively, a two-way ANOVA can be used (with the usual assumptions of normality and homogeneity of variances among aggregate samples). We start by stating that the reproducibility variance is the sum of three terms: within-laboratory variance (σ_0^2), between-laboratory variance (σ_L^2), and laboratory×sample interaction (σ_{LS}^2). With ANOVA, repeatability and reproducibility standard deviations can be expressed in terms of their mean squares. For repeatability, this is simply the square root of the within-laboratory variance:

$$s_r = \sqrt{MS_{\text{error}}} \tag{4}$$

Reproducibility is calculated as (13)

$$s_R = \sqrt{\frac{1}{kr}(MS_{\text{lab}} - MS_{\text{lab×sample}}) + \frac{1}{r}(MS_{\text{lab×sample}} - MS_{\text{error}}) + MS_{\text{error}}} \tag{5}$$

Constants k and r are the number of different aggregate samples and number of replicates, respectively. When reproducibility is based on one duplicate, equation 5 simplifies to

$$s_R = \sqrt{\frac{1}{2}(MS_{\text{lab}} + MS_{\text{error}})} \tag{6}$$

Both repeatability and reproducibility standard deviations may be reported in dimensionless units by dividing each by its corresponding mean (across all laboratories), which when represented as a percent is known as the coefficient of variation (CV) or relative standard deviation (RSD). Pictorially, repeatability and reproducibility can be explained by a two-sample chart, in which in lieu of the customary plotting of readings from two close, independent laboratory samples the readings for duplicates of the same aggregate sample across many laboratories are plotted in an X-Y plot (Fig. 2).

The diagonal line represents zero repeatability, such that the perpendicular distances of the points from this line collectively represent the actual repeatability. Dispersion in the direction parallel to the diagonal line represents lab-to-lab variation, which for small within-laboratory error approximates the reproducibility error. More often than not, the intent of the collaborative study is to quantify the variation across laboratories (i.e., reproducibility) rather than within a laboratory (i.e., repeatability). This type of plot can also be used to visually describe precision using two close, but slightly different, laboratory samples, instead of duplicates (discussed later).

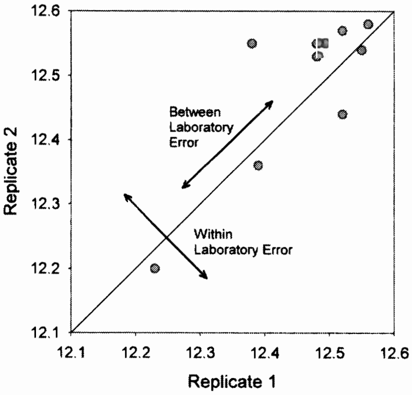


Fig. 2. Two-sample chart.

Treatment of Outliers

As often happens, the study director will be faced with the issue of removing test portion measurements or entire laboratories from the analysis of precision. Despite the tests for outliers that we describe below, it is the ultimate responsibility and judgment of the study director whether to remove a measurement or entire laboratory from the analysis. Assuming that all collaborators' data have been collected properly, an initial comparison of the collaborators can be performed by assigning ranks to the collaborators (1 to n) for each test portion measurement and then summing each collaborator's ranks over the number of independent samples studied. A table is created that lists the measurements for each independent sample across all collaborators along with its rank with respect to all collaborators (Fig. 3). Collaborators with tie values are each assigned an average value for the ranks they would otherwise occupy if the values were close but not equal.

By summing each laboratory's ranks over all independent samples, it is possible to determine whether one or more laboratories are consistently low or high with respect to the other laboratories. Ranks may be compared with confidence levels that provide the upper and lower limits within which the ranks lie at a given level of probability (e.g., 95%). In the example of the near-infrared prediction of wheat protein content shown in Fig. 3, we see that laboratory A, with a summed rank of 5.0, fell outside the limits of 10 and 45 that correspond to a 95% confidence interval (10,11).

IUPAC has adopted a systematic set of procedures for outlier detection that has been adopted by AOACI (2). Briefly, a Cochran test is performed to determine the consistency in levels of within-laboratory variability of each aggregate sample across all laboratories (Fig. 4). A laboratory whose replicate-sample measurements differ by a statistically significant larger value (typically applied as a one-tail test at $P = 0.025$) is considered a Cochran outlier. The Cochran test statistic is calculated as

$$\text{Cochran test statistic} = 100\frac{s^2_{\text{replicate max}}}{\sum_{i=1}^{n} s^2_{\text{replicate } i}} \tag{7}$$

The variables $s^2_{\text{replicate } i}$ and $s^2_{\text{replicate max}}$ represent the variance of test portion measurements from laboratory i and the largest within-laboratory variance over the n laboratories, respectively. Cutoff values for the Cochran maximum variance ratio (one-tail, 2.5% rejection level) are tabulated in AOACI's guidelines (2).

The second set of outlier tests, know as Grubbs tests, examines the laboratories for unusually high or low values for each aggregate sample (Fig. 5). Calculation of the Grubbs statistic involves determining the standard deviation of test portion average measurements across all laboratories and the standard deviation of the remaining laboratories when the laboratory with the lowest average is removed (s_L). Likewise, the standard deviation when the laboratory with the highest average is removed (s_H) is also calculated. The Grubbs statistic be-

Lab	For Each Sample: Protein Content (Rank)					Sum of Ranks
	a	b	c	d	e	
A	8.58 (1.0)	10.95 (1.0)	11.48 (1.0)	11.54 (1.0)	12.00 (1.0)	5.0*
B	8.93 (3.0)	11.25 (9.5)	11.59 (5.5)	11.81 (4.5)	12.09 (3.0)	25.5
C	8.84 (2.0)	11.02 (2.0)	11.58 (4.0)	11.89 (8.0)	12.06 (2.0)	18.0
D	9.02 (8.0)	11.08 (3.5)	11.74 (8.0)	11.90 (9.0)	12.18 (4.0)	32.5
E	8.98 (6.0)	11.12 (5.0)	11.56 (3.0)	11.81 (4.5)	12.26 (6.0)	24.5
F	9.01 (7.0)	11.13 (6.0)	11.55 (2.0)	11.69 (2.0)	12.27 (8.0)	25.0
G	8.97 (5.0)	11.14 (7.0)	11.80 (9.0)	11.82 (6.0)	12.27 (8.0)	35.0
H	9.11 (10.0)	11.25 (9.5)	11.59 (5.5)	11.97 (10.0)	12.27 (8.0)	43.0
I	8.94 (4.0)	11.24 (8.0)	11.83 (10.0)	11.86 (7.0)	12.25 (5.0)	34.0
J	9.09 (9.0)	11.08 (3.5)	11.69 (7.0)	11.77 (3.0)	12.28 (10.0)	32.5

Fig. 3. Table of collaborators' measurements and corresponding ranks for five aggregate samples (near-infrared transmittance predictions of protein content in bulk wheat). Data from Delwiche and coworkers (4).

comes the percent decrease in standard deviation units when the laboratory with the most deviant average is removed:

$$\text{Grubbs test statistic} = \text{larger of}\left[100\left(1-\frac{s_L}{s}\right), 100\left(1-\frac{s_H}{s}\right)\right] \quad (8)$$

As with the Cochran outlier test, values for the single-value Grubbs outlier (two-tail, 2.5% rejection level) are tabulated using AOACI's guidelines (2). The Grubbs test is also applied to three additional cases: removal of the two lowest averages, removal of the two highest averages (as opposed to the situation described in equation 8, in which the single lowest or single highest laboratory is removed), and removal of the lowest and highest laboratories simultaneously. The exact sequence of testing for outliers is shown in a flowchart (Fig. 6). IUPAC established the rule that the removal of outliers should cease if the fraction of removed laboratories exceeds 2 of 9 of the starting number of laboratories. Further, the number of retained laboratories should not fall below eight. Finally, the decision to remove an outlying laboratory must be done after careful consideration by the study director upon review of the collaborator's notes. In addition to statistical outliers, measurements may also be removed for an identifiable reason, such as a lab deviating from the method, a sample arriving in poor condition, a mistake in labeling or recording of the data, etc. Data such as these are considered invalid and should not be included in the statistical analysis.

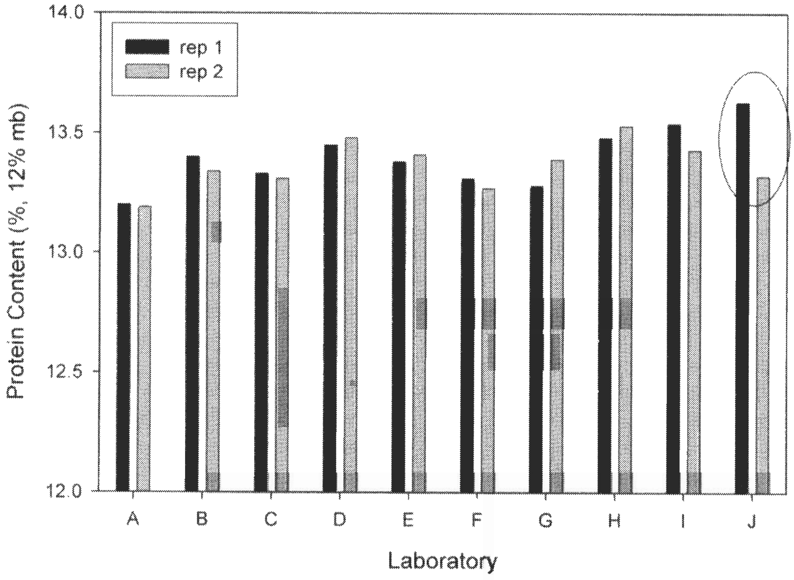


Fig. 4. Example of a Cochran outlier.

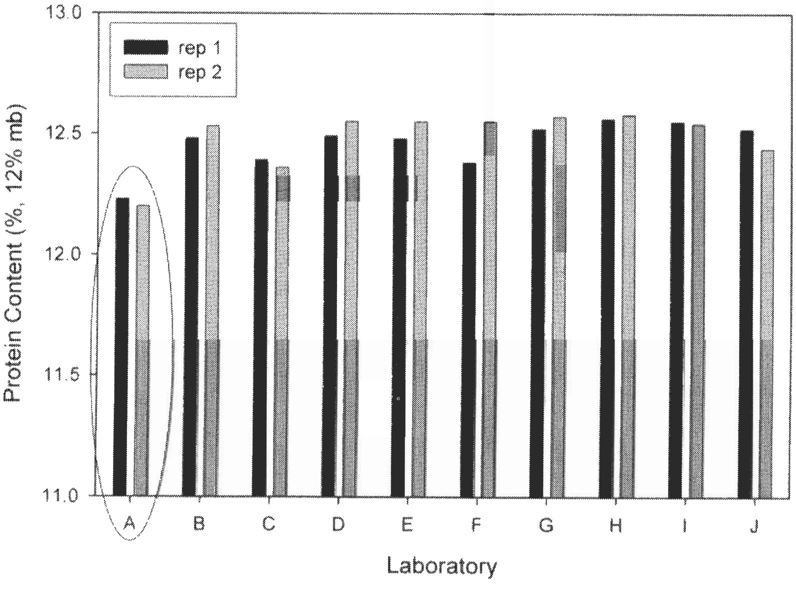


Fig. 5. Example of a Grubbs outlier.

Repeatability Without Blind Duplicates: Youden Matched-Pairs

To guard against the tendency of analysts to innocently censor their data on blind duplicates by negating errant readings and substituting them with additional analyses, the late William Youden, then of the National Bureau of Standards, suggested that repeatability could be measured through the analysis of two close, but slightly different, consignments (although still considered as one material). We will call the consignments x and y, and the measurements from laboratory i will be x_i and y_i, respectively (12). One motive for this approach is seen in the two-sample plot (also known as a Youden plot) shown in Fig. 7.

The construction of this plot differs from that of Fig. 2 in that the 45-degree line, rather than passing through the origin, is drawn through the coordinate point that corresponds to the x and y means of the plotted data. By requiring that the laboratory samples of a matched pair are close in value (AOACI guidelines specify <5% in actual difference), the measured difference between the two results will reflect only the within-laboratory (repeatability) random error and the true difference between the laboratory samples and not include any systematic error associated with the laboratory. Therefore, as with the case of duplicates, the dispersion of the points in the direction perpendicular to the 45-degree line characterizes the repeatability of the method.

In recent years, McClure (7) has characterized the statistical design of Youden matched-pairs as that of a repeated measures model and has identified the necessary assumptions in the Youden procedure: 1) the reproducibility variances of x and y (s^2_{Rx} and s^2_{Ry}) are equivalent, 2) there is no interaction between x and y, and 3) the estimates of method precision are valid only for the material under study (i.e., the test sample factor is a fixed effect). With these assumptions, the expressions for repeatability and reproducibility, respectively, become

$$s_r = \sqrt{\frac{1}{2(n-1)}\sum_{i=1}^{n}(d_i - \bar{d})^2} \quad (9)$$

and

$$s_R = \sqrt{(s^2_{Rx} + s^2_{Ry})/2} \quad (10)$$

where $s^2_{Rx} = [1/(L-1)][\sum x_i^2 - (\sum x_i)^2/L]$ and $s^2_{Ry} = [1/(L-1)][\sum y_i^2 - (\sum y_i)^2/L]$. A t test at $L - 2$ degrees of freedom (and a typical significance level of $\alpha = 0.05$) is required for the reproducibility variances, s^2_{Rx} and s^2_{Ry}, to ensure compliance with McClure's first assumption (Part 6, Appendix D in AOACI *OMA Program Manual* [3]), in which the following formula is used to calculate t:

$$t = \frac{(s^2_{Rx} - s^2_{Ry})(L-2)^{1/2}}{2[(s^2_{Rx})(s^2_{Ry}) - (\mathrm{cov}_{xy})^2]^{1/2}} \quad (11)$$

where $\mathrm{cov}_{xy} = [1/(L-1)][\Sigma x_i y_i - (\Sigma x_i \Sigma y_i)/L]$.

Written in the form of an ANOVA

$$s_r = \sqrt{MS_{\text{error}}} \qquad (12)$$

and

$$s_R = \sqrt{\frac{1}{2}(MS_{\text{lab}} + MS_{\text{error}})} \qquad (13)$$

A Spreadsheet Approach to Precision Calculations

Identifying outliers and calculating precision parameters by hand can be a daunting task for the nonstatistician. AOACI recognized this, and in the early 1990s provided spreadsheet programs designed to work with LOTUS 123 and early versions of Excel to perform the calculations. Although the spreadsheets were functional and simple to use, the software macro language they employed eventually became obsolete. These early spreadsheets have since been replaced with updated Excel versions complete with macros that enhance the user–program interface. There are two versions of the spreadsheets, one for blind replicates and the other for Youden matched-pairs. In accordance with the current AOACI guidelines, the spreadsheets calculate method performance statistics separately for each material (aggregate sample) studied.

The appearance of the program for calculation of blind replicates is shown in Fig. 8. Absent from this figure are the interface windows that guide the user through the process. Upon opening the program, the user is prompted to enter the analyte and aggregate sample ID. The number of replicates (single, duplicate, or triplicate), the units of measurement, appropriate decimal format, and recovery (if applicable) are then entered. Next, the data are entered, the HORRAT value (not discussed in this article) and collaborative study statistics are calculated, and outliers are identified. If outliers are detected, the user manually deletes the data from the offending laboratory(ies), and the statistics are recalculated.

In the program for Youden matched-pairs, additional calculations are done to test for equivalence of variance (equation 11) and to determine whether the difference in analyte concentration exceeds 5% between the members of the pair. Data from up to 100 laboratories can be entered, but currently, outlier identification is available for only up to 75. Although the spreadsheet identifies statistical outliers, the decision to remove the data remains in the hands of the user.

These programs were developed by J. M. Lynch with the intention of making the calculations as easy and robust as possible. They are available free of charge upon completion of a licensing agreement (to obtain copies, contact Joanna M. Lynch at JL72@cornell.edu).

General Use of ANOVA to Calculate Precision

ANOVA is routinely used as a statistical analysis tool in experimentation to examine treatment differences. ANOVA compares between-treatment variation to within-treatment variation to determine statistically significant differences between treatments. ANOVA can be used to determine if an analytical method is performing equally well for all aggregate samples, varieties, or treatments being tested in a collaborative study. In accordance with IUPAC harmonized guidelines for collaborative study procedures, the previously discussed spreadsheet program essentially uses a one-way (single-factor) ANOVA to determine method performance characteristics for a single material. In the following, we present the equivalent analysis using the shorthand format of ANOVA tables.

In the single-factor ANOVA model table for a blind duplicates design (Table I), L = the number of participating labs, and r = the number of replications of a single (i.e., laboratory) sample or treatment sent to each lab. In the blind duplicates design, r = 2, and the degrees of freedom for within laboratories becomes simply L, and the total degrees of freedom becomes $2L - 1$. The performance parameter calculations of repeatability and reproducibility are obtained from information provided in the ANOVA table, such that the repeatability variance (s_r^2) is MS_{wl} and the reproducibility variance (s_R^2) is $(MS_{bl} + MS_{wl})/2$.

In the single-factor ANOVA model table for a Youden matched-pair design (Table II), the within-laboratory variance term from the blind duplicates model is split into two components, a replicate effect and a labo-

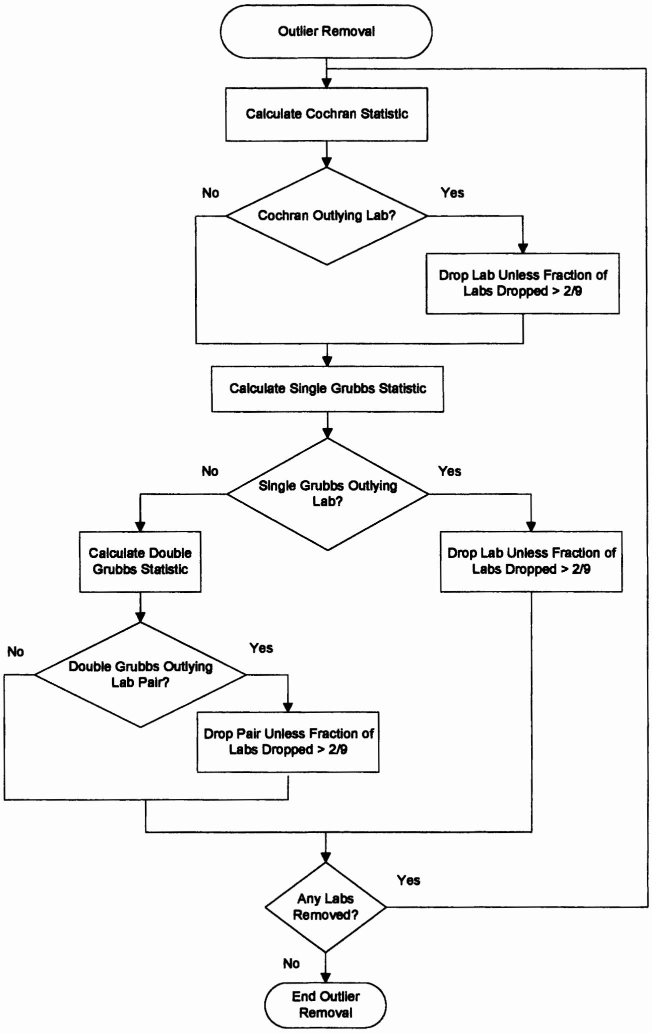

Outlier Removal
Calculate Cochran Statistic
No
Cochran Outlying Lab?
Yes
Drop Lab Unless Fraction of Labs Dropped > 2/9
Calculate Single Grubbs Statistic
No
Single Grubbs Outlying Lab?
Yes
Calculate Double Grubbs Statistic
Drop Lab Unless Fraction of Labs Dropped > 2/9
No
Double Grubbs Outlying Lab Pair?
Yes
Drop Pair Unless Fraction of Labs Dropped > 2/9
Yes
Any Labs Removed?
No
End Outlier Removal

Fig. 6. Flowchart of outlier removal procedure as adapted by IUPAC (2).

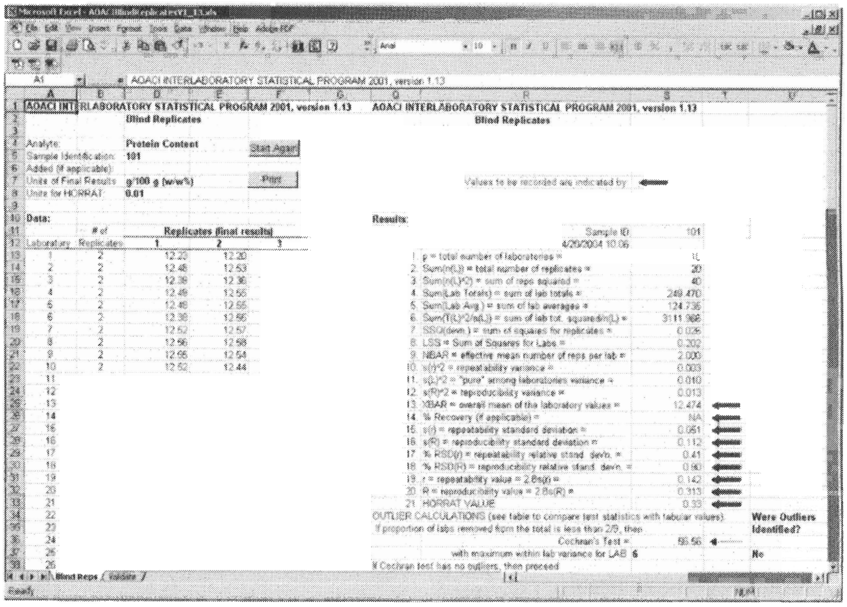
Fig. 7. Two-sample plot for a Youden matched-pair.

ratory×replicate interaction term. With the replicate effect removed from the blind duplicates within-laboratory variance term, the laboratory×replicate effect becomes the within-laboratory variance term for the Youden matched-pair model. The replicates in a Youden matched-pair design are not exact duplicates of each other as in the blind duplicates design. There is error associated with differences between replicates that is removed from the within-laboratory variance component for a Youden matched-pair model. For this model, the value of r is 2, because the replicates are still paired; the degrees of freedom for replicates is 1, and the degrees of freedom for laboratory×replicate is $L - 1$. The performance parameter estimates for a Youden matched-pair design ANOVA are calculated exactly the same way as for a blind duplicates design. The only difference is that the replicate effect has been removed from the within-laboratory variance term used in the Youden matched-pair calculations.

As an example, assume nine laboratories are analyzing two replicates of a treated powder for a mineral detection method collaborative study. The first case considers the two laboratory samples to be blind duplicates (Table III), and the second case considers them to be Youden matched-pairs (Table IV). The calculated performance parameters between cases are not very different in this example, but depending on the accuracy required for acceptance of a method, meaningful differences are possible.

The distinctions between the designs become more apparent when treatment differences are incorporated into an analysis. In collaborative studies, it is valuable to evaluate a method over a variety of materials for greater method applicability. We use the AOACI term "material" and the statistical term "treatment" interchangeably. Treatments could be a different matrix, such as wheat, rye, or corn flour. Treatments could also be distinct levels of a variable, such as high protein content or low protein content in wheat flour alone.

Researchers examine performance parameters for a particular treatment but often are also interested in analyzing treatment differences and ascertaining how the method performs overall. Factorial ANOVAs can be used to calculate overall treatment performance parameters of a method, as well as to determine treatment differences. The study director will need to check with the board or committee overseeing collaborative studies (be it within AACC, AOACI, ISO, or others) to determine whether simultaneous analysis of more than one material is allowed. Due to inherent ANOVA assumptions, statistical tests on the homogeneity of variance among materials must be performed prior to conducting an ANOVA. Software packages incorporate many types of statistical tests for variance homogeneity, such as Hartley's F-max, Bartlett's, and Levene's tests (8,9).



A1 AOACI INTERLABORATORY STATISTICAL PROGRAM 2001, version 1.13

AOACI INTERLABORATORY STATISTICAL PROGRAM 2001, version 1.13
Blind Replicates

Analyte:	Protein Content
Sample Identification:	101
Added (if applicable):	
Units of Final Results	g/100 g (w/w%)
Units for HORRAT	0.01

Data:

Laboratory	# of Replicates	Replicates (final results) 1	2	3
1	2	12.23	12.20	
2	2	12.48	12.53	
3	2	12.39	12.36	
4	2	12.49	12.55	
5	2	12.48	12.55	
6	2	12.38	12.56	
7	2	12.62	12.57	
8	2	12.56	12.58	
9	2	12.55	12.54	
10	2	12.52	12.44	
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23				
24				
25				
26				

Values to be recorded are indicated by ⟵

Results:

	Sample ID	101
	4/20/2004 10:06	
1.	p = total number of laboratories =	10
2.	Sum(n(L)) = total number of replicates =	20
3.	Sum(n(L)^2) = sum of reps squared =	40
4.	Sum(Lab Totals) = sum of lab totals =	249.470
5.	Sum(Lab Avg.) = sum of lab averages =	124.736
6.	Sum(T(L)^2/n(L)) = sum of lab tot. squared/n(L) =	3111.966
7.	SSQ(devn.) = sum of squares for replicates =	0.026
8.	LSS = Sum of Squares for Labs =	0.202
9.	NBAR = effective mean number of reps per lab =	2.000
10.	s(r)^2 = repeatability variance =	0.003
11.	s(L)^2 = "pure" among laboratories variance =	0.010
12.	s(R)^2 = reproducibility variance =	0.013
13.	XBAR = overall mean of the laboratory values =	12.474 ⟵
14.	% Recovery (if applicable) =	NA ⟵
15.	s(r) = repeatability standard deviation =	0.051 ⟵
16.	s(R) = reproducibility standard deviation =	0.112 ⟵
17.	% RSD(r) = repeatability relative stand. devn. =	0.41 ⟵
18.	% RSD(R) = reproducibility relative stand. devn. =	0.90 ⟵
19.	r = repeatability value = 2.8s(r) =	0.142 ⟵
20.	R = reproducibility value = 2.8s(R) =	0.313 ⟵
21.	HORRAT VALUE	0.33 ⟵

OUTLIER CALCULATIONS (see table to compare test statistics with tabular values) — Were Outliers Identified?
If proportion of labs removed from the total is less than 2/9, then
Cochran's Test = 56.56 ⟵
with maximum within lab variance for LAB 6 — No
If Cochran test has no outliers, then proceed

Microsoft Excel - AOAC BlindReplicatesV1_13.xls
File Edit View Insert Format Tools Data Window Help Adobe PDF
Arial
10
Start Again
Print
Blind Reps
Validate
Ready
NUM

Fig. 8. Spreadsheet program for determination of repeatability and reproducibility and outlier testing.

Table I. Single-factor ANOVA model table for a blind duplicates design

Source	Degrees of Freedom	Sum of Squares	Mean Squares
Between labs	$(L - 1)$	SS_{bl}	$MS_{bl} = SS_{bl}/(L - 1)$
Within labs	$L(r - 1)$	SS_{wl}	$MS_{wl} = SS_{wl}/L(r - 1)$
Total	$Lr - 1$		

Table II. Single-factor ANOVA model table for a Youden matched-pair design

Source	Degrees of Freedom	Sum of Squares	Mean Squares
Between labs	$(L - 1)$	SS_{bl}	$MS_{bl} = SS_{bl}/(L - 1)$
Replicate	$(r - 1)$	SS_r	$MS_R = SS_R/(r - 1)$
Lab×Rep	$(L - 1)(r - 1)$	SS_{wl}	$MS_{wl} = SS_{wl}/(L - 1)(r - 1)$
Total	$Lr - 1$		

Table III. ANOVA model table for blind duplicates design with two replicates

Source	Degrees of Freedom	Mean Squares	
Between labs	8	35.69	
Within labs	9	11.32	
Total	17		Overall mean = 37.07 (from ANOVA printout)

$s_r^2 = MS_{wl} = 11.32$	$s_r = 3.36$	$\text{RSD}_r = 100(3.36/37.07) = 9.06\%$
$s_R^2 = (MS_{bl} + MS_{wl})/2 = 23.5$	$s_R = 4.85$	$\text{RSD}_R = 100(4.85/37.07) = 13.08\%$

Table IV. ANOVA model table for Youden matched-pair design with two replicates

Source	Degrees of Freedom	Mean Squares	
Between labs	8	35.69	
Replicate	1		
Within labs	8	10.68	
Total	17		Overall mean = 37.07 (from ANOVA printout)

$s_r^2 = MS_{wl} = 10.68$	$s_r = 3.27$	$\text{RSD}_r = 100(3.27/37.07) = 8.82\%$
$s_R^2 = (MS_{bl} + MS_{wl})/2 = 23.2$	$s_R = 4.82$	$\text{RSD}_R = 100(3.27/37.07) = 13.0\%$

As a practical and statistical matter, laboratories in collaborative studies are not always given blind duplicates if treatments are to be compared, but instead are given suitably matched replicates (same treatment, different batches). This situation corresponds to an extension of a Youden matched-pair design for more than two replicates. In the factorial ANOVA table incorporating treatments (Table V), L = the number of participating labs, t = the number of treatments (materials) sent to each lab, and r = the number of replicates per treatment sent to each lab. The performance parameter estimates for a factorial ANOVA are calculated exactly the same way as for the single-factor Youden matched-pair design. The only difference is in how the within-laboratory variance is partitioned. The factorial ANOVA within-laboratory variance term is laboratory×replicate(treatment).

Table V. Factorial ANOVA table incorporating treatments

Source	Degrees of Freedom	Mean Squares
Between labs	$(L-1)$	MS_{bl}
Treatment	$(t-1)$	
Lab×Trt	$(L-1)(t-1)$	
Replicate(Trt)	$t(r-1)$	
Lab×Rep(Trt)	$t(r-1)(L-1)$	MS_{wl}
Total	$Lrt-1$	

As mentioned earlier, this type of factorial ANOVA is a repeated-measures design, where the replicates are nested within treatments. Although the replicates within a treatment are made from different batches, each laboratory is sent the same replicates. This reduces the possibility that treatment or method differences may be mainly due to differences between the replicates. Because replicates within a treatment are not exact duplicates, replicate error is partitioned out of the within-laboratory variance component in the same way as a single-factor ANOVA for Youden matched-pairs. Further complications for factorial ANOVA arise when one considers whether the treatments and laboratories are fixed effects, random effects, or a combination of effects (a mixed-effects model). There are differences in F-test denominators used for testing treatment or laboratory differences for each of the different effects models, but the performance parameter calculations are not affected by whether the model is a fixed-, random-, or mixed-effects model.

Table VI shows the results for the same nine laboratories used in the previous example, with two replicates per treatment per lab and an added powder formulation treatment. The repeatability standard deviation, s_r, doesn't differ much from the analysis of each powder treatment separately, implying that the method is being performed consistently for different powders within a

Table VI. Factorial ANOVA table with two replicates per treatment and an added treatment

Source	Degrees of Freedom	Mean Squares
Between labs	8	169.28
Treatment	1	
Lab×Trt	8	
Replicate(Trt)	2	
Lab×Rep(Trt)	16	13.96
Total	35	

Overall mean = 59.7 (from ANOVA printout)

$s_r^2 = MS_{wl} = 13.96$	$s_r = 3.74$	$\text{RSD}_r = 100(3.74/59.7) = 6.26\%$
$s_R^2 = (MS_{bl} + MS_{wl})/2 = 91.62$	$s_R = 9.57$	$\text{RSD}_R = 100(9.57/59.7) = 16.0\%$

laboratory. The reproducibility standard deviation, s_R, is larger than the analysis for a single powder treatment. This may indicate there is a laboratory×treatment interaction, where one or more laboratories are performing the method differently for each powder treatment. This can be examined directly through the factorial ANOVA F tests.

There are both advantages and disadvantages to using factorial ANOVA for performance parameter calculations. The advantages are that researchers can test varietal, sample, and treatment differences within the same factorial analysis used to generate the repeatability and reproducibility estimates in a collaborative study. In most collaborative studies, there is interest in the scope of the method over as diverse a test set as possible.

Using a factorial ANOVA to obtain performance parameter estimates gives a detailed statistical evaluation of a method's overall applicability. Examining laboratory× treatment interactions to determine whether one or more laboratories are having problems with a particular treatment for methodological inconsistencies can identify outlier laboratories. One analysis is sufficient to determine treatment differences and calculate performance parameters instead of a separate analysis for each treatment.

One disadvantage of a factorial ANOVA occurs when performance parameters are desired for each variety or treatment. Other disadvantages include making sure the correct model is used and, in the case of blind replicates, the added computational complications when replicates are not balanced. If the goal of a collaborative study is to show the individual effectiveness of a method when a wide variety of treatments or samples are tested, then use of the one-way ANOVA for each sample is recommended. If the goal is to show the scope or overall effectiveness of a method over a wide variety of treatments, use of a factorial ANOVA is warranted.

Conclusions

Unified analytical procedures play a vital role in today's global cereals and cereal products market by assuring that accurate information on nutritional value and functional properties is shared among growers, processors, and consumers. The underpinning of these procedures is the collaborative, or interlaboratory, study. Although often lacking in glamour, because by its nature a collaborative study procedure is often the verification of a conservative, pretested method (i.e., the thrill of discovery is absent), the importance of such a study cannot be overemphasized.

The AACC Approved Methods Committee stands as the association's gatekeeper on cereals methodology. Through this primer, we hope that cereal chemists, engineers, and quality-control specialists can gain knowledge and appreciation of AACC's Approved Methods and the implied research behind each approved method.

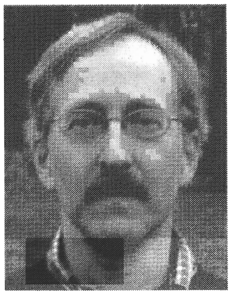
Acknowledgments

We thank T. Nelsen (USDA-ARS. Peoria, IL) for helpful comments on the review of this paper.

References

1. AACC. *Approved Methods of the American Association of Cereal Chemists, 10th ed.* The Association, St. Paul, MN, 2000.
2. AOAC. Guidelines for collaborative study procedure to validate characteristics of a method of analysis. J. AOAC Int. 78(5):143A. 1995.
3. AOACI. *OMA Program Manual.* Published online at www.aoac.org/vmeth/omamanual/omamanual.htm. The Association, Gaithersburg, MD. 2003.
4. Delwiche, S. R., Pierce, R. O., Chung, O. K., and Seabourn, B. W. Protein content of wheat by near-infrared spectroscopy of whole grain: Collaborative study. J. AOAC Int. 81:587, 1998.
5. IUPAC. Protocol for the design, conduct and interpretation of collaborative studies. Pure Appl. Chem. 60:855, 1988.
6. IUPAC. Nomenclature for sampling in analytical chemistry. Pure Appl. Chem. 62:1193, 1990.
7. McClure, F. D. A statistical evaluation of the Youden matched-pairs procedure. J. AOAC Int. 82:375, 1999.
8. Neter, J., Wasserman, W., and Kutner, M. H. *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs.* 3rd ed. Richard D. Irwin, Inc., Homewood, IL, pp. 614-624, 1990.
9. Palmquist, D. My view. Weed Sci. 45:745, 1997.
10. Thompson, W. A., and Willke, T. A. On an extreme rank sum test for outliers. Biometrika 50:375, 1963.
11. Wernimont, G. T. *Use of Statistics to Develop and Evaluate Analytical Methods.* W. Spendley, ed. Association of Official Analytical Chemists, Arlington, VA, p. 183, 1985.
12. Youden, W. J. The collaborative test. J. Assoc. Off. Anal. Chem. 46:55, 1963.
13. Youden, W. J., and Steiner, E. H. *Statistical Manual of the AOAC.* Association of Official Analytical Chemists, Arlington, VA, 1975.

The Authors



Stephen R. Delwiche is a research scientist with the USDA/ARS in Beltsville, MD, where he has worked since 1990. He holds B.S. and Ph.D. degrees from Cornell University (Ithaca, NY) and a M.S. degree from North Carolina State University (Raleigh), all in agricultural and biological engineering. His current assignment focuses on development of optical equipment, techniques, and multivariate statistical models for quality and food safety concerns in small grains. He serves as an associate editor for *Cereal Chemistry*, cochair of the AACC Near Infrared Analysis Technical Committee, past chair of the AACC Engineering and Processing Division, and the AACC/AOAC INTERNATIONAL liaison officer.

Debra E. Palmquist worked as a research mathematician/statistician for USDA/ARS in Reno, NV, from 1979 to 1999. She received B.S. and M.S. degrees in mathematics from the University of California, Riverside, and University of Nevada, Reno, respectively. She is currently the ARS Midwest Area biometrician (Peoria, IL). She is also a consulting statistical editor for *Weed Science*, AACC Approved Methods Committee statistician, member of the AACC Check Sample Committee, and chair of the AACC Statistical Advisory Technical Committee.

Joanna M. Lynch studied at Cornell University (Ithaca, NY), where she received a B.S. degree (with distinction) in clinical and biochemical nutrition and an M.S. degree in clinical nutrition. She joined the Food Science Department at Cornell in 1986 and has worked in various research and laboratory management positions in the field of dairy chemistry. Lynch has been involved with the design and implementation of numerous collaborative studies, resulting in 11 AOAC Official Methods. As a Fellow of AOAC INTERNATIONAL, she has served in a variety of capacities, including chair of the Committee on Commodity Foods and Commodity Products and member of the Official Methods Board. She developed the spreadsheets used for calculating method performance statistics.