• *Technical*

Accuracy in the Control Laboratory

T. F. WATERS, HARRY SMITH JR., R. C. STILLMAN,
The Procter and Gamble Company, Ivorydale, Ohio

WHEN OUR FIRST visitor from space is welcomed in the U.S.A., one of the features of his tour will be the inspection of some typical factories on which our civilization is based. Whether impressed or not, he will find that, despite the vast differences he sees in the various factories, there are several activities common to all and one of them will be an area called the control laboratory, usually tucked away in a corner of the top floor of one of the remote buildings. What are we prepared to answer when he asks for an explanation of the word "control," used so often in describing the laboratory? Do we really mean that the analytical results obtained in the control laboratory will be the major determinants in regulating the subsequent operation of the factory in its self-imposed job of producing identical units?

Consideration of these and other probable questions soon leads to the proper definition of the function of the control laboratory. We must accept the realization that our scale of production would be impossible if we had to obtain analytical evidence of what we had done before deciding what next to do. Mass production of goods which will adhere to a common tight set of specifications is only feasible when we follow repetitively standardized operations with standardized raw materials. We have not yet reached the level of competence wherein automation with control laboratory information fed back to really control operation is a reality. So our control laboratory today has the basic responsibility for testing samples taken from the process and for verifying after the event that the process was proceeding normally. It is equally true that if abnormal or so-called "out-of-control" results are found, the analytical information is utilized to determine what to do to restore control. The frequency of such occurrences must be minimized to obtain economical operation, so such situations should seldom occur.

Even though most of the analytical data we obtain from our control labs are to verify that we were doing our intended task properly rather than to determine what we must do now, the analytical data must be accurate. Since we simply cannot afford to test completely all of our finished product before shipment, we rely on the evidence that our process control was effective and know from this that our finished product will be found to perform as expected. Obviously there can be false assurance that the process is proceeding normally. Therefore a problem constantly besetting the laboratory administrator is the improvement of the accuracy of every analytical determination made in the control laboratory. It may be of interest to report on the system used in Procter and Gamble and recommend therefrom for similar programs used on a larger scale.

OUR PROGRAM consists of sending samples representing the various important analytical techniques periodically to the control laboratories. The frequency varies from monthly for the commoner techniques to semi-annually for less important techniques. Samples cover the range of techniques from wet chemical through instrumental and include such subjective judgments as odor or taste. The samples are prepared carefully to be uniform and are chosen to cover the range of the specifications for which the analysis is utilized, with additions occasionally to the production material so that the resultant sample is one that would signal real trouble when analyzed properly in routine operation. As some samples change with age, analyses are made on a schedule determined by our ability to send the samples to the most remote participating laboratory. Results are collated, and rankings are published monthly for all labs.

As most programs do, this one began as an emergency when a disagreement between two laboratories on an analysis caused an interchange of samples to determine which result was correct. In retrospect, the wisdom of such a program on a routine basis was recognized. Thus about 40 years ago we began to send out check samples to our labs periodically. The number and frequency of the samples grew, until by 1932 enough laboratories were participating and enough samples were being run each month that a ranking system was begun. The growth has continued until at last count 36 laboratories received an average of 54 samples monthly with a maximum of 67. We attempt to have samples available on which sufficient determinations are made to equal 2% of our normal monthly analytical load. The relative standings are calculated monthly, both on the last month's samples and on a six- and twelve-month basis. The factory management follows the results with interest, and we really understand the full intensity of this interest only when a factory finds that it has been awarded a lower position than it has earned as a result of a mistake that we have made in compiling the results.

The basic purpose of the program is to improve the accuracy with which all analyses are performed so every effort is made to ensure that the conditions under which the samples are run duplicate the conditions under which routine determinations are made. To achieve this goal, a specific set of rules has evolved. The more important are these. Only one analysis shall be made, and the result must be reported. No special checking of the accuracy of the analysis shall be made when the unknown sample is run, such as running a known sample in parallel, unless the same check method is used routinely for all samples. The analyst regularly making the determination shall run the special sample, and if there is more than one analyst making the determination routinely, the samples are assigned so that all share equally over a period of time. The analyses are made on the same days of the month and must be reported within a specific period. Any discussion of results or questions shall be directed only to our office. The laboratory manager must attest that these conditions have been met when he submits the results.

ONCE the laboratories have run the special samples (called cooperative samples) and reported the results, there is the equally important scoring of the results and calculating of the rankings achieved. The scoring system is based on the following thinking. It is an accepted fact that any analytical result is only an estimate of the true value of the characteristic in question. Repetitive analyses of identical material will yield results that will be distributed about a central value, which will be the true value only if bias is absent from the measurement but from a practical standpoint can be considered to be the true value. With most processes acceptable results are obtained over a significant range of operating conditions; and as the inherent sampling and analytical error is usually small compared with the acceptable inherent-process variation, periodic single analytical results usually are satisfactory for control purposes. Thus the expected variation in an analytical result from the true value usually is not important, and any specific figure reported within the range to be expected around the true value is useful to operating personnel in confirming that the process remains in control or, if outside that range, in indicating the need for corrective action. However an analytical result that falls outside of the expected range around the true value through error gives false information and can cause wrong managerial action or at least a delay until a recheck is completed. In our program, as in all such programs, the true value of the analysis is not known, and the most likely estimate of the true value is used. This is the average of all results that appear to be free of gross error.

The between-laboratory standard deviation for the analysis is used to rate the results. This measure of the test precision has been calculated normally only after at least 100 results are available, using all the results except those obviously in error. The calculation is as follows. The mean value for each set of samples for the analysis in question is calculated. The difference between each individual result and the mean value of the set to which it belongs is calculated. The pooled standard deviation is determined from

$$s = \sqrt{\frac{\sum d^2}{n-x}}$$

where d = differences, n the number of results, and x the number of sets of samples. Thus if 20 laboratories each run the same five samples, n = 100 and x = 5. We assume then the s so determined equals the true standard deviation σ.

In some analyses the σ is not uniform over the range of the characteristic but varies in size, depending on the magnitude of the characteristic. In such a case larger σ's are used for successive ranges of the characteristic or a continuous function is established. Of course, a pooled standard deviation cannot be calculated in this case, but pooling can be done for sets of samples with essentially common variance. The standard deviations are recalculated periodically, when about 100 results have accumulated since the previous calculation, to test whether or not the σ has changed.

The actual calculation of the rating for a laboratory is as follows: For each set of samples the median result is determined. Results more than 3.2 σ from the median are identified and excluded from determination of the mean value. The mean of the remaining results is calculated. All results are scored against the mean on the following basis:

TABLE I

Distance from final mean value	Penalty points	Identification
$\pm 0 - 0.8\ \sigma$	0	0 underscore
$\pm 0.81 - 1.6\ \sigma$	1	1 underscore
$\pm 1.61 - 2.4\ \sigma$	2	2 underscores
$\pm 2.41 - 3.2\ \sigma$	6	3 underscores
$> \pm 3.2\ \sigma$	24	4 underscores–circled

If the number of underscores is found to exceed twice the number of reporting laboratories on any sample, while such a result is perfectly possible statistically, previous experience has indicated that a nonuniform sample has been distributed so the results are discarded. This happens to less than 2% of the samples. When all samples have been rated, the laboratory rating is calculated from the formula:

$$\text{Rating} = 100 - \frac{\text{total penalties} \times 17}{\text{total determinations}}$$

The factor 17 is arbitrarily chosen to make a rating of about 90% apply to a laboratory which attains the theoretically normal distribution of errors for our own managerial reasons. The use of 0.8 σ intervals in the above table rather than the usual unit σ intervals is merely to obtain better separation of the rankings. You will note the increasing weight or nonlinearity of the penalty assigned as the reported result is farther away from the accepted result for reasons discussed earlier.

THE RATING of a laboratory which achieved this statistically normal distribution of errors would be calculated as shown in Table II, and for comparison the over-all ratings of our participating laboratories for the years 1953 and 1958 are also shown.

TABLE II

Class	Normal distribution	1953	1958
	%		
A ± 0 to 0.8 σ from mean	57.62	71.00	70.93
B ± 0.81 to 1.6 σ from mean	31.42	20.45	21.16
C ± 1.61 to 2.4 σ from mean	9.32	5.21	4.96
D ± 2.41 to 3.2 σ from mean	1.50	1.42	1.64
E $> \pm 3.2\ \sigma$ from mean	0.14	1.92	1.31
Rating	89.4	85.4	87.7

It is evident that the performance is not statistically normal as performance is, better than expected, close to the average, and poorer for the less precise results. We believe that the better-than-expected precision is a combination of achieving some degree of true repetitiveness in the analysis and of the unavoidable special care that the sample receives despite the restrictions discussed earlier. The poorer-than-expected precision at the bottom is simply evidence of gross errors in manipulation, transcribing, calculation, reporting, etc. The extent of skewing is not great however as very close to half of the laboratories achieve a rating better than 89.4 with the top lab at 96.91 for 1958.

There is more than you might expect to the improvement in precision from 1953 to 1958 that Table II indicates, for major reductions in the magnitudes of various analytical σ's have been made in the period,

while the number of laboratories reporting has about doubled.

Examples illustrating the reductions achieved in the σ's as improved precision has been achieved are shown in Table III for 10 analyses selected at random.

TABLE III

Analysis	1953 σ	1958 σ	% Reduction
Moisture (granules)	.26	.19	27
Moisture (paste)	1.00	.68	32
Titration SO_3	.10	.056	44
N_2	.10	.078	22
Hydroxyl value	2.20	1.74	21
Periodate glycerine	.312	.242	22
Lye Bé	.14	.14	....
Iodine value	.36	.36	....
Saponification value	1.0	0.80	20

The mere publication of the ranking list monthly provides what we believe to be a healthy stimulus in the form of the manager's questions to the laboratory head about his plans to improve his relative standing.

In addition to the cooperative analysis program which has been described, each laboratory carries on its own internal auditing program. Basic in such a plan is knowledge of the within-laboratory analytical standard deviations, and these should be known in each laboratory for all the important analyses. For any analysis the within-lab σ should be smaller than and usually no more than two-thirds of the between-lab σ. If it is found not to be so, an investigation is made to uncover the cause of the poor precision. Comparisons of the σ's achieved by various analysts on the same job will indicate where further training is needed. Two methods of establishing the internal σ's are used.

a) Following a regular schedule, a number (usually 2–4%) of samples that have been run in routine work are resubmitted as new samples without the analyst's knowledge and the σ calculated by the method of duplicates:

$$s = \sqrt{\frac{\Sigma (R_1 - R_2)^2}{2N}} = \sigma$$

b) A sample for which the analytical value is known is analyzed repetitively and the σ determined, using the deviations from the known value (used often in training).

$$s = \sqrt{\frac{\Sigma d^2}{N}} = \sigma$$

A routine comparison of the between- and the within-lab σ's is made as shown in Table IV and within-lab σ's are underlined to indicate where the laboratory should make an investigation to improve its performance on the analysis in question.

WHILE the foregoing has been developed to fit the needs of control laboratories, it has long been recognized that commercial laboratories have the same need of ensuring as high a degree of accuracy as feasible for all of their routine analyses, and it is believed that the same principles should be useful in designing a uniform method for comparing accuracy of an industry's laboratories.

TABLE IV

Analysis	Between-lab. σ	Within-lab. σ				
		Lab. A	B	C	D	E
Bé on silicate	.10	.06	.058	.06	.07	.033
Color (glycerine)	.16	.09	.13	.10	.08	.09
Ppm (chlorophyll)	.005	.003	.003	.004	.004	....
Glycerine (C.P.)	.24	.24	.12	.19	.13	.20
Glycerine on fats	.13	.11	.06	.10	.10	.06
Iodine value	.36	.25	.27	.23	.13	.14
H_2O—hot plate	.06	.03	.04	.05	.04	.03
Na_2O—lye	.124	.04	.03	.05	.05	.09
SO_3 titration	.10	.07	.04	.08	.10	.02

Much has been accomplished in improvement of laboratory accuracy and precision in our industry over the years by the excellent work of the Smalley Committee of the American Oil Chemists' Society, but it is noted that no one series of samples is handled as to scoring like any other series so this review is presented in the belief that a uniform method of scoring is possible and would make the work even more effective. It should be said that the A.O.C.S. industry program seems already to be well ahead of similar programs of other industries as far as we can determine.

When the various Smalley series are compared against each other for their use of the statistical principles important in such work, we find little uniformity. Table V lists these principles and their application as we understand them in each case, and we shall discuss them and make recommendations for a common method.

TABLE V

Smalley Committee—Industry Analytical Programs

	No. analyses per sample	Exclusion of results	Rating criterion	Diff. RC for diff. range	Linear penalty
Vegetable oils	1	No	Fixed	Yes	Yes
Edible fats	1	"Irwin's Criteria"	Between σ	Yes	Yes
Inedible fats	1	"Irwin's Criteria"	Between σ	Yes	Yes
Glycerine	2	"Obvious"	Within σ	Yes	Nearly
Oil seeds	1	"Outlying"	Fixed	Yes	Yes
Oil seed meals	1	"Obvious"	Fixed [a]	No	Yes

[a] From median.

Number of Analyses per Sample. Since we believe it to be basic that any program for promoting and rating laboratory accuracy should reproduce as closely as possible the handling of the routine work of the laboratory, the same number of analyses per sample should be made as is needed with the routine samples. Likewise if multiple runs are made with routine samples, should we not attempt to make repetitive runs unnecessary for economic reasons? Therefore we recommend one analysis per sample unless there is a compelling reason for more.

Exclusion of Results. To be really useful we believe ratings must be calculated from the best estimate available of the true value of the sample. Thus all results which are really in error should be identified and excluded from the calculations of the mean value to be used as the best estimate. These are the results

which, if available and accepted as the sole measure of a sample as in routine work, would lead to wrong decisions.

There are many statistical criteria for identifying such outliers as these results to be excluded are termed. Without going more fully into the reasons, we shall state only that our choice between them is a function of underlying assumptions, availability of probability tables, simplicity of handling, and sample size. Two categories are needed, the first for methods of analysis where the analytical error (σ) to be expected is known, and the second where the σ is unknown. The recommendation is as follows:

TABLE VI

Standard deviation	Sample size	Criterion
Known	1–20	Studentized range
Known	>20	Normal deviate
Unknown	1–25	Grubb's
Unknown	>25	Normal approximation

Short descriptions and references to these criteria follow.

Studentized Range (1): This criterion assumes sampling from a normal distribution. The statistic calculated is

$$q = \frac{w}{\sigma} \quad \text{where } w = \text{range}$$

The value is referred to table of the w/σ distribution for the desired α risk level, and the most extreme observation from the median is deemed to be an outlier as long as the calculated q exceeds the tabled q_α. The tables (2) do not go above a sample size of 20.

Normal Deviate: This criterion assumes a random sample from a normal population of unknown mean but known σ. Any observation which falls beyond 3.2 σ from the median is rejected as an outlier. This is the method which we have used with good results. Another good criterion for this class is Irwin's (3), which is already used by two of the Smalley series, but we prefer the Normal Deviate because of its greater ease of calculation.

Grubb's (4): These criteria assume a sample of size n from a normal distribution of unknown σ. For testing the largest observation the statistic is calculated

$$\frac{s_n^2}{s^2} = \frac{\sum_{i=1}^{n-1} (x_i - \bar{x}_n)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

where

$$\bar{x}_n = \frac{1}{n-1} \sum_{i=1}^{n-1} x_i \quad \text{and} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Other calculations are used for the two largest observations or other combinations. The probability distributions of these statistics are tabled for sample sizes of 1 to 25, and results are regarded as outliers if the tabled figures are exceeded.

Normal Approximation: This is the same method as used for the larger sized sample where σ is known. Here s (which is used as the estimate of σ) is calculated from the data of the experiment itself, and the proper value of "t" from the Student "t" table is used. Depending on the degrees of freedom, the range will be 3.2–3.7 s.

In the above discussion of rejection of outliers, the σ that is used is the standard deviation of a single observation per laboratory. Thus if we are to reject outliers in a set of single results furnished by several laboratories, then the between laboratory error should be used.

Rating Criterion. After the best estimate of the true value of the sample has been obtained, what criterion should be used to rate the results reported? We prefer a statistically derived measure of the analytical variability to any arbitrarily chosen rating-criterion. If a statistical measure is to be used, should it be the within or the between laboratory σ? Everyone will agree that it would be ideal if the two σ's were equal and the need for the choice did not exist, but at present there exist biases between laboratories that cause the between lab σ to be larger than the true variability of the analysis itself.

We believe further that we really are less interested in how closely a laboratory can check itself than we are in how closely a laboratory can approach the accepted analytical value for the sample in question.

Thus we recommend the use of the between-laboratory σ for ranking performance of laboratories, but for determination of the inherent analytical variability of a method as may be needed in methods development work or analytical precision investigations of a given laboratory, the within-lab σ should, of course, be used. A report has been presented by the Statistics Committee of the A.O.C.S., which outlines fully the recommended method of determining both within- and between-lab σ's to be used when analytical procedures are investigated by A.O.C.S. committees. This method of necessity requires the running of multiple determinations, but there is no conflict with our earlier recommendation for a single analysis per sample as the present recommendation is only for the purpose of rating laboratories after methods which are used have been accepted.

Different Ranking Criterion for Different Range of Sample Value. If the σ of the analysis changes significantly as the magnitude of the analytical result changes, then the need is obvious for the proper σ relative to the range of the given sample.

Linear or Nonlinear Penalty. We recommend the use of nonlinear penalties from the reasoning discussed more fully above. While an analytical result that varies a small amount from the true value may be a perfectly useful result, a result that causes the wrong managerial action to be taken is harmful and should thus be penalized much more heavily. It might be argued that this principle, while right for the control laboratory, is not right for the commercial laboratory where the ability to determine exactly the true value of the sample should be stressed; but we believe, even in analytical results governing commercial transactions, that the average over a period of time of acceptable analytical results will approach the true average value, and it is the avoidance of real errors that is important.

We offer these recommendations in an effort to make even more successful the long industry cooperative effort in improvement and simplification of our analytical procedures.

REFERENCES

1. Tippett, L. H. C., "The Extreme Individuals and the Range of Samples Taken from a Normal Population," Biometrika, vol. 17 (1925).
2. Dixon and Massey, "Introduction to Statistical Analysis," Table 8A, pp. 405, McGraw Hill Book Company Inc., 1957.
3. Irwin, J. O., "On a Criterion for the Rejection of Outlying Observation," Biometrika, vol. 17 (1925).
4. Grubbs, F. E., "Sample Criteria for Testing Outlying Observations," Annals of Mathematical Statistics, vol. 21, No. 1, March 1950.

[Received May 4, 1959]

Effect of Composition and Polymorphic Form on the Hardness of Fats[1]

R. O. FEUGE and WILMA A. GUICE, Southern Regional Research Laboratory,[2] New Orleans, Louisiana

HARDNESS is an important consideration in the performance of confectionery fats. Ordinarily fats are desired which are relatively hard and brittle at room temperature yet soften and melt at slightly higher temperatures. Conceivably a measurement of hardness also can be used to determine whether or not a fat-containing confection has been tempered properly (2).

The hardness of confectionery fats, which may contain 80% or more of solids at room temperature, should not be regarded as being identical with the consistency of plastic, semisolid fats like shortening and margarine oil, which generally contain less than 20% solids at room temperature. The general property of hardness has been variously defined as resistance to local penetration, scratching, cutting, wear or abrasion, and yielding. The multiplicity of definitions indicates that hardness is not a fundamental property but rather a composite one including yield strength, work hardening, true tensile strength, and modulus of elasticity.

On the assumption that a mass of fat crystals resembles in certain important respects a mass of metal crystals, it might be expected that a modification of the Brinell test for metals should be well suited for measuring the hardness of solid or substantially solid fats. Apparently tests bearing any resemblance, even remotely, to the Brinell test for metals have been used very infrequently with fats. Ravich and Volnova (3) applied such a test to tristearin-tripalmitin and stearic acid-palmitic acid mixtures. Von Rosenberg (4) described a test procedure for fats and waxes which embodied some of the principles of the Brinell test. Recently in our laboratory an instrument and test procedure were devised and found to be satisfactory in testing fats and waxes (2).

In our modification of the Brinell test a perfectly round steel ball having a diameter as small as 0.1250 in. or as large as 0.5000 in. is pressed for 1 min. with a force of 0.2 to 6.0 kg. into the surface of the test specimen. The applied force is selected so that the diameter of the impression ranges between 15 and 45% of the diameter of the ball. The hardness index is calculated from the formula:

$$H = \frac{P\,(100)}{\frac{\pi D}{2}\,(D - \sqrt{D^2 - d^2})}$$

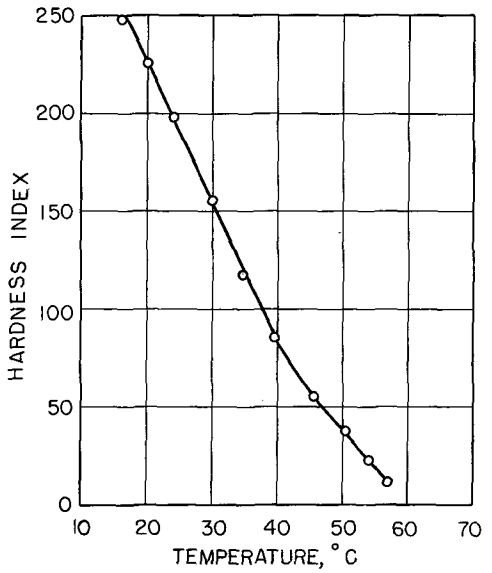

FIG. 1. Hardness curve for completely hydrogenated cottonseed oil melted and heated to 90°C., solidified at 26°C., and stored for several months at 26°C.

where H is the hardness index, P is the weight on the ball in kilograms, D is the diameter of the ball in millimeters, and d is the diameter of the impression in millimeters. The denominator of the above equation represents the curved area of the impression, while the factor 100 in the numerator reduces the dimensions of the hardness index to kilograms per square centimeter. The index is practically independent of ball size and test load if the other test conditions are confined to certain ranges (2).

This communication presents data on the effect of composition and polymorphic form on the hardness of fats. It should provide new information useful in the production of better fat products and also should provide a background for the evaluation of any new test data obtained with the instrument and technique.

Temperature Effects

Temperature has a marked influence on the hardness of a solid fat, even when polymorphic transformations, changes in crystal size, and partial melting are not involved. The decrease in hardness as the temperature increases is relatively gradual, even for a pure triglyceride, and is not an abrupt phenomenon like the melting of a pure compound. In fact, over the temperature range at which fats are commonly utilized

[1] Presented at the 50th Annual Meeting, American Oil Chemists' Society, New Orleans, La., April 20-22, 1959.
[2] One of the laboratories of the Southern Utilization Research and Development Division, Agricultural Research Service, U. S. Department of Agriculture.