(Reprinted from INDUSTRIAL QUALITY CONTROL, Vol. XV, No. 11, May 1959)

Graphical Diagnosis of Interlaboratory Test Results

W. J. YOUDEN
National Bureau of Standards, Washington, D. C.

Introduction

Interlaboratory or round robin programs to evaluate the performance of test procedures will always be with us. New materials require new tests. New, and hopefully better, test procedures are developed for old products. Test procedures are used to ascertain whether a product meets the specification set down for the product. A double problem confronts the producer. There is bound to be a certain amount of variation in his product. And there is bound to be variation in the test results made on a given sample of the product. The impact of the errors of measurement associated with the test procedure is obvious because half the tests made on a product that just meets specification will rate the product below specification.

Test Procedures and Production Costs

It is customary to manufacture purposely a product that exceeds specification in order to allow for testing errors. The larger these testing errors, the greater the excess quality that must be built into the product to insure the acceptance of nearly all lots that are in fact equal to or better than the specification. The manufacturer already has to contend with variation in the process. Considerable saving in manufacturing costs can be affected by reducing the margin between the quality level set for production and that called for in the specification. The savings attainable with improved test procedures are a strong inducement for the improvement of test procedures. Interlaboratory test programs of varying degrees of thoroughness are frequently used to establish the performance of existing procedures.

Missed Opportunities in Interlaboratory Test Programs

Strangely enough modern statistical tests such as the analysis of multifactor studies and the isolation of components of variance have not made the contribution expected of them. Part of this no doubt comes about because these more sophisticated statistical techniques are not too well understood by some of those in the laboratories that run the tests. It is all very well for someone with statistical skill to set up an intricate interlaboratory test program and analyse the data but this still leaves the problem of interpreting the statistical jargon to those directly concerned. Even when this interpretation is undertaken the report is apt to read somewhat along these lines. "Duplicates run by the same operator in the same laboratory show excellent agreement. Agreement between different operators in the same laboratories is not quite so good, and very poor between results from different laboratories. Results on different days do not agree as well as those obtained on the same day." This is a brief summary of the interpretation that is made after the statistical analysis shows that practically all the F-tests are significant. Unhappily almost all concerned were already aware of the state of affairs just described and want to know what can be done to improve matters. It is just here that statisticians have not risen to the opportunities presented by interlaboratory test programs.

When all is said and done, what we want is rather simple. We want to know whether the test procedure as set forth is capable of yielding acceptable agreement among results from different laboratories. If the results are not acceptable, we would like some specific indication of what is wrong with the procedure. If the procedure appears to be reasonably good but there are some disturbing discrepancies, we would like to know which laboratories are having trouble and if possible why they are having trouble. And most important we should be able to get this information back to the laboratories concerned in such a form that the diagnosis is believed. For only so will these laboratories take any action to correct the difficulties.

Graphical Representation of Results

The graphical procedure is based upon a very simple interlaboratory program. Samples of two different materials, A and B, are sent to a number of laboratories which are asked to make one test on each material. The two materials should be similar and be reasonably close in the magnitude of the property evaluated. This will avoid complications that may arise from differential behavior of the two test materials. A second pair of samples are circulated at a later time if there are only a few participating laboratories. The pairs of results that are reported by the laboratories are used to prepare a graph.

The graph is prepared by drawing the customary x-axis at the bottom of the paper and laying off on this axis a scale that covers the range of results for material A. At the left the y-axis is provided with a scale in the same units that includes the range of results reported for material B. The pair of results reported by a laboratory are then used to plot a point. There will be as many points as there are reporting laboratories. After the points are plotted a horizontal median line is drawn parallel to the x-axis so that there are as many points above the line as there are below it. A second median line is drawn parallel to the y-axis and so placed that there are as many points on the left as there are on the right of this line. Figure 1 shows the seven-day tensile strengths reported by 25 laboratories on two cement samples. Two of the laboratories are so patently separated from the other 23 that they are not used in determining the position of the median lines.

Diagnosis of the Configuration of Points

The two median lines divide the graph paper into four quadrants. In the ideal situation where *only* random errors of precision operate the points are expected to be equally numerous in all quadrants. This follows because plus and minus errors should be equally likely. In any existing test procedure that has come to my attention the points tend to concentrate in the upper right and lower left quadrants. This means that laboratories tend to get high results on both materials or low results on both materials. Here is evidence of individual laboratory biases. There is evidence of this state of affairs in Fig. 1. The more pronounced this tendency to individual bias the greater the departure from the expected circular distribution of points about the intersection of the median lines.

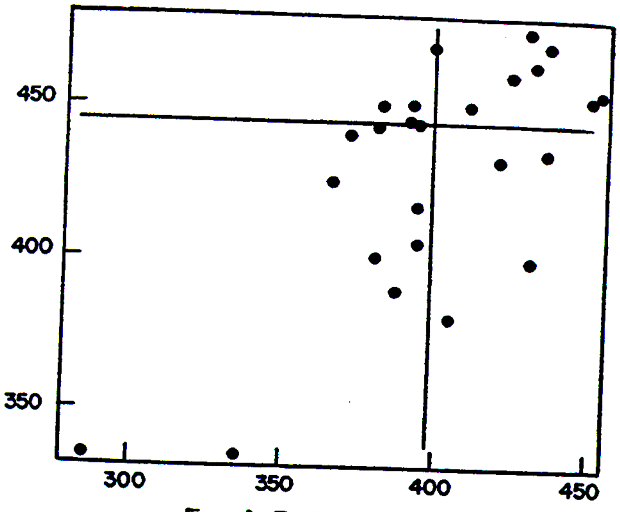

Figure 1—Tensile Strength

Figure 2 shows 15 points plotted from phthalic anhydride determinations on two paint samples. The points tend to scatter more or less closely along a line approximately bisecting the upper right and lower left quadrants. There is reason to expect the line to make a 45 degree angle with the axes when the same scale is used for both axes and the two materials are sufficiently similar so that the dispersion of the results is about the same for each material.

A test procedure that yields results like those in Fig. 2 is probably in need of more careful description. In its present form the procedure apparently is open to individual modifications that do have an effect upon the results. The procedure rather than the laboratories should be considered as a possible source of the difficulty even though the difficulty is exhibited by a large scatter among the results from the different laboratories. When the points lie closely along the 45 degree line the conclusion may be drawn that many of the laboratories

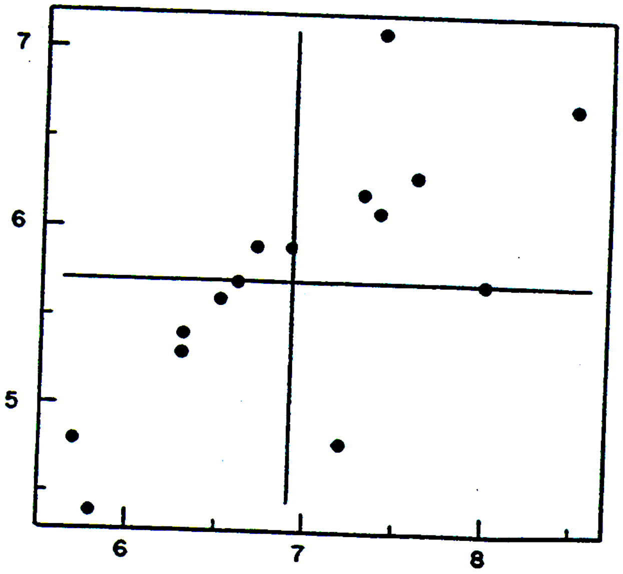

Figure 2—Percent Phthalic Anhydride

are following rather carefully their own versions of the test procedure.

Checking on Sample Variation

There is no possibility of the distribution of points in Fig. 2 arising from lack of uniformity among the samples distributed from each material. If the stock is heterogeneous, some samples will be high, some low, and this will be true for both materials. The pairs of samples distributed to the laboratories will be of four kinds:

high in A, high in B
high in A, low in B
low in A, high in B
low in A, low in B

The four possible combinations have the same probability of occurrence and would result in the test results being nearly equally divided among the four quadrants. Concentration of the points in two quadrants rules out questions of sampling heterogeneity.

On the other hand if there is a roughly circular distribution of points but with a disappointingly widespread scatter, the diagram does not reveal whether this arises from sampling difficulties or poor precision of the test results. If sampling is considered a possible source of difficulty the following modification in the assignment of samples should be tried. If there are 2N laboratories, prepare N *double-size* samples for each material. Carefully mix and divide each double-size sample into two usual size samples.

Double size		Samples	
sample	Laboratory	A	B
1	1	1A	1B
	2	1A′	1B′
2	3	2A	2B
	4	2A′	2B′
N	2N−1	NA	NB
	2N	NA′	NB′

The samples are assigned to laboratories as shown above. It should be possible to mix and divide each double-size sample into two closely matching regular samples. These samples are assigned to a pair of laboratories. If there are sampling difficulties the plotted points should tend to occur in doublets. Two laboratories getting the two carefully mixed halves should check each other and have their points close together. This involves a little extra work in getting out the samples and no extra work for the participating laboratories. If the points corresponding to the two halves of a double-size sample are separated as much, on the average, as points from different double samples, the dispersion cannot be ascribed to sampling. In addition to noting the spacial distribution the projections of the points on the axes may also be used to see whether just one of the materials was heterogeneous.

Interpretation of Out-of-Line Results

So far the large aspects of the diagram have been examined. The individual points can now be considered and in particular those points most distant from the intersection of the median lines. Almost always one or more points are so far out of the picture that it is better not to compress the scale in order to show them. Such points should be ignored in locating the median lines. (See Fig. 1.) The more distant points tend to fall into one or the other of two categories. Either the point is far out and remote from both axes or far out and fairly close to one or the other axis. In the latter case, the result is fairly good on one material and very bad on the other. Examples of such points are found in Figures 2 and 3. Often the explanation is simple—a mistake in typing, or calculation, or some simple blunder that sometimes can be corrected by going back to the records. If the same laboratory shows up in such a manner on succeeding pairs of materials, this implies carelessness on the part of the laboratory. The laboratory can do good work but often does not. Occasionally a laboratory has difficulty with one material and not with the other but this is not likely to occur with similar materials.

Points in the upper right or lower left quadrants that are far removed from the intersection of the median lines and that are not near either axis reflect a tendency to get either high results on both materials or low results on both materials. There are examples in all the figures. The more consistent a laboratory is in its work the more likely its point will lie in the proximity of the 45 degree line. A point far out along this line suggests the possibility that the laboratory concerned has introduced some modification into the test procedure. A laboratory finding itself in this situation should check carefully the prescribed procedure for performing the test and endeavor to locate the cause of the large bias.

All of the above interpretation can be made while keeping anonymous the identity of the plotted points. When circulating a report of the interlaboratory test it might be helpful to circle in red the point belonging to the laboratory in the copy going to that laboratory. That would save the laboratory from consulting its files to locate itself and would display prominently just where the laboratory stood in reference to the whole group. This vivid picturing of a laboratory's position should stimulate the laboratory to some self examination that could hardly avoid having beneficial results.

Estimating the Precision of the Test Procedure

The above discussion does not exhaust the information to be gleaned from this graphical representation. Assuming that the two materials are similar in type and nearly equal in magnitude for the property the dispersion among the results reported for A should be about the same as the dispersion of the B results. In that event the 45 degree line through the intersection of the medians makes possible an estimate of the *precision* of the data. Often an interlaboratory test undertakes to differentiate among the laboratories in respect to precision. Not only does this require large numbers of measurements from each laboratory but differences in precision usually turn out to be unimportant in comparison with bias errors and careless errors. No violence at this stage seems to be done by assuming about the same precision for all the laboratories.

The perpendicular distance from each point to the 45 degree line can be used to form an estimate of the precision. The estimate of the standard deviation of a *single* result is obtained by multiplying the average length of the perpendiculars by $\sqrt{\pi/2}$ or 1.2533. These perpendiculars need not be measured on the graph paper. Instead, write d wn for each laboratory the difference (A−B) keeping track of the signs. Call these differences $d_1, d_2, \ldots d_n$. Calculate $\bar{d}$, the algebraic average difference. Subtract $\bar{d}$ from each difference and obtain a set of corrected differences $d_1', d_2', \ldots d_n'$. The average of the absolute values of these differences when multiplied by $\sqrt{\pi}/2$ or 0.886 gives an estimate of the standard deviation.

TABLE I—Data and Calculations on Percent Insoluble Residue in Cement Reported by 29 Laboratories

Laboratory	Percent Residue A	Percent Residue B	A — B	(A—B) — 0.095
1	0.31	0.22	0.09	—0.005
2	0.08	0.12	—0.04	—0.135
3	0.24	0.14	0.10	0.005
4	0.14	0.07	0.07	—0.025
5	0.52	0.37		
6	0.38	0.19	0.19	0.095
7	0.22	0.14	0.08	—0.015
8	0.46	0.23		
9	0.26	0.05	0.21	0.115
10	0.28	0.14	0.14	0.045
11	0.10	0.18	—0.08	—0.175
12	0.20	0.09	0.11	0.015
13	0.26	0.10	0.16	0.065
14	0.28	0.14	0.14	0.045
15	0.25	0.13	0.12	0.025
16	0.25	0.11	0.14	0.045
17	0.26	0.17	0.09	—0.005
18	0.26	0.18	0.08	—0.015
19	0.12	0.05	0.07	—0.025
20	0.29	0.14	0.15	0.055
21	0.22	0.11	0.11	0.015
22	0.13	0.10	0.03	—0.065
23	0.56	0.42		
24	0.30	0.30	0.00	—0.095
25	0.24	0.06	0.18	0.085
26	0.25	0.35		
27	0.24	0.09	0.15	0.055
28	0.28	0.23	0.05	—0.045
29	0.14	0.10	0.04	—0.055
Average	0.229	0.134	0.095	0.053

The data on percent insoluble residues reported by 29 laboratories are given in Table I and plotted in Fig. 3. There are three points far out along the 45 degree line and one far out on the y-axis. These laboratories were excluded from the calculations shown in Table I. The last column shows the differences between the two results diminished by the difference between the two sample averages. The average, 0.053, shown at the bottom of this column is the average of the *absolute* values, i.e., ignoring the signs. Multiplying 0.053 by 0.886 gives 0.047 as the estimate for the standard deviation of a single result. Probably this is inflated by leaving in the two laboratories turning in the very low results for sample A.

This estimate of the standard deviation for precision leads to the construction of circles (centered on the intersection of the median lines) within which any given percentage of the points can be expected to fall should the laboratories be able to eliminate all bias or constant errors. The multiples of the standard deviation that include various percents of the points are given in Table II.

Thus a circle whose radius is about 2.5 to 3.0 times the standard deviation gives a fair idea of the smallest circle that could be expected to contain nearly all points after the elimination of the constant errors that are causing the points to congregate in the upper left and lower right quadrants. Generally a fair number of points will lie outside such a circle. The laboratories responsible for these points almost certainly have somehow got substantial systematic errors incorporated in their techniques. Multiplying the standard deviation obtained above by 2.45 gives the radius of the circle that should include 95 percent of the laboratories if individual constant errors could be eliminated. This circle is drawn in Fig. 3. Seven further laboratories are outside the circle including the two who got the benefit of the doubt and were retained in the computation. This examination has directed attention to at least six of the laboratories that might well go over their method of making this determination of insoluble residue.

TABLE II—Probability Table for Circular Normal Distribution

Percent of the Points Within Circle	Multiple b of the Standard Deviation
10	0.459
20	0.668
25	0.759
30	0.845
40	1.011
50	1.177
60	1.350
70	1.552
75	1.665
80	1.794
90	2.146
95	2.448
99	3.035

Note: Percent = 100[1 — exp(—b^2/2)]

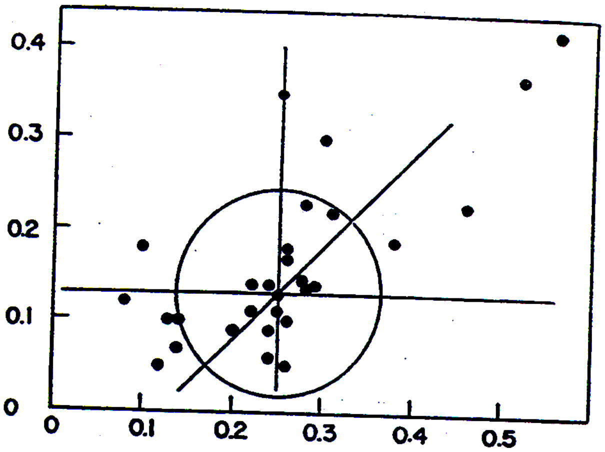

Figure 3—Percent of Insoluble Residue

If the number of laboratories in the program is rather small, the way to accumulate more points is to send the laboratories additional pairs of samples from different materials. A chart is prepared for each pair of materials and the median lines drawn in. The charts are now superimposed so that the points of intersection of the median lines coincide and, of course, the median lines also. All points are then transferred to one sheet of paper with one pair of median lines. As there are only a few laboratories each can be assigned an identifying symbol.

Figure 4 shows the reports made by eight laboratories determining CaO in cement. The laboratories are

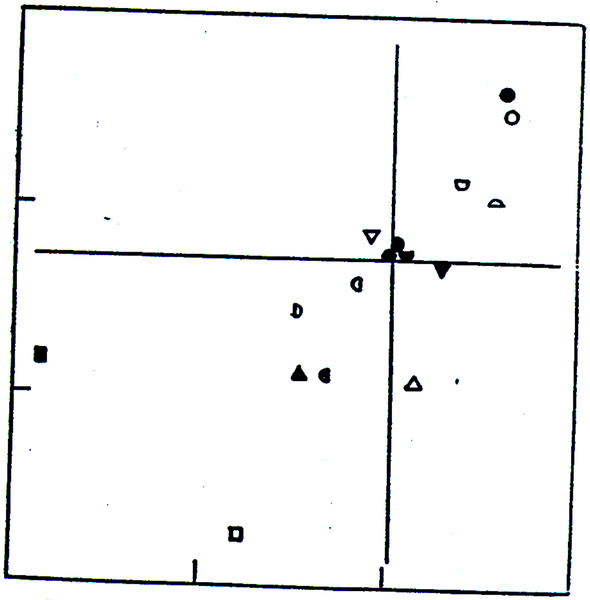

Figure 4—CaO in Cement (interval equals one percent)

identified by symbols. The hollow symbols show the results for the *first pair* of samples. The corresponding solid symbols show the work of these same laboratories on a *second pair* of samples. Few as these data are they serve to indicate the things that we want to know about the test procedure and about the laboratories. Clearly this procedure is one that is vulnerable to individual bias. Two of the eight laboratories appear in the same region for both pairs. The circle laboratory is very consistent—and gets the highest results. The square laboratory gets very low results and is not very precise as shown by the fact that the two squares are separated by a much greater distance than any of the other seven pairs. Using this chart some possibly helpful suggestions could be passed along.

Discussion

The two materials used in this double-sample program were specified to be similar in type and in the magnitude of the property measured. Sometimes the measurement errors are proportional to the magnitude under measurement and this will show up in a greater scatter of the points along one of the axes. Particulaт types of samples may give trouble in just some of the laboratories. The thorough study of a test method must include consideration of these possible complications. Naturally a more comprehensive interlaboratory test program will be required to explore these aspects of the test procedure. A thorough study in one laboratory usually reveals these complications.

Summary of Advantages of Graphical Diagnosis

The double-sample, graphic analysis scheme described in this article offers a number of advantages.

(1) An unusually light burden is imposed on each laboratory

(2) The graphical procedure greatly facilitates presentation of the results in a convincing manner

(3) No statistical background is required to follow the reasoning and no computations are required to demonstrate the general presence of constant errors and the gross deviations of individual laboratories

(4) A minimum of computation is imposed upon the individual collating the results

(5) The use of a circle of 2.5 or 3.0 σ radius shows the individual laboratories whether or not their method of carrying out the test has in some way become saddled with a substantial constant error

(6) Most important the direction for improvement is clearly indicated

a. A long, narrow ellipse directs attention to a more careful description of the procedure or even to the need for modification

b. Wild points far out near either axis indicate erratic work

c. Wild points far out along the 45 degree line are strong evidence of substantial deviations from the specified procedure

d. General prevalence of constant errors is indicated by a substantial proportion of the points lying outside the 2.5 σ circle

Experience has already indicated that a certain few laboratories are found too frequently in the most distant positions from the intersection of the median. Improved performance from these few laboratories may go far to restore confidence in a test procedure. There is no substitute for careful work in the laboratory.